IV.B. Toward a Theory of Group Estimation*

NORMAN C. DALKEY

Introduction

The term "Delphi" has been extended in recent years to cover a wide variety of types of group interaction. Many of these are exemplified in the present volume. It is difficult to find clear common features for this rather fuzzy set. Some characteristics that appear to be more or less general are: (1) the exercise involves a group; (2) the goal of the exercise is information; i.e., the exercise is an inquiry, (3) the information being sought is uncertain in the minds of the group; (4) some preformulated systematic procedure is followed in obtaining the group output.

This vague characterization at least rules out group therapy sessions (not inquiries), team design of state-of-the-art equipment (subject matter not uncertain), brainstorming (procedure not systematic), and opinion polls (responses are not treated as judgments, but as self-reports). However, the characterization is not sufficiently sharp to permit general conclusions, e.g., concerning the effectiveness of types of aggregation procedures.

Rather than trying to deal with this wide range of activities, the present essay is restricted to a narrow subset. The subject to be examined is group estimationthe use of a group of knowledgeable individuals to arrive at an estimate of an uncertain quantity. The quantity will be assumed to be a physical entity—a date, a cost, a probability of an event, a performance level of an untested piece of equipment, and the like.

Another kind of estimation, namely, the identification and assessment of value structures (goals, objectives, etc.) has been studied to some extent, and a relevant exercise is described in Chapter VI. Owing to the difficulty of specifying objective criteria for the performance of a group on this task, it is not considered in the present paper.

To specify the group estimation process a little more sharply, we consider a group $I = \{ I_i \}$ of individuals, an event space $E = \{ E_j \}$ where E can be either discrete or continuous, and a response space $R = \{ R_{ij} \}$ which consists of an estimate for each event by each member of the group. In addition, there is an external process $P=\{ P (E_j) \}$, which determines the alternatives in E which will occur. Depending on the problem, P can either be a δ (delta)-function on E— i.e., a specification of which event will occur-or a probability distribution P_j on the event space. In general P_j is unknown. For some formulations of the group estimation process, it is necessary to refer to the a priori probability of an event. This is not the same as the external process, but rather, is (in the present context) the probability that is ascribed to an event without knowing the individual or group estimates. This a priori probability will be designated by $U = \{ U (E_j) \}$.

* Research reported herein was conducted under Contract Number F30602-72-C-0429 with the Advanced Research Projects Agency, Department of Defense.

In many cases the R_{ij} are simply selections from E. The weatherman says, "It will rain tomorrow"-a selection from the two-event space rain tomorrow and no rain tomorrow. The long-range technological forecaster says, "Controlled nuclear fusion will be demonstrated as feasible by 1983"—selection of a single date out of a continuum. In these cases the R_{ij} can be considered as 0's and 1's, 1 for the selected event and 0 for the others. Usually, the 0's are left implicit. More complex selections can be dealt with – " I t will either rain or snow tomorrow." "Controlled nuclear fusion will be demonstrated in the interval 1980-1985"by allowing several 1's and interpreting these as an or-combination. Selections can also be considered as special cases of probability distributions over the event space. In the case of probability estimates, the R_{ij} can be probability assignments for discrete alternatives, or continuous distributions for continuous quantities.

A kind of estimate which is sometimes used in applied exercises, but which is not directly expressible in terms of elementary event spaces, is the estimation of the functional relationship between two or more variables (e.g., the extrapolation of a trend). Such an estimate can be included in the present formalism if the relationship is sufficiently well known beforehand so that all that is required is specification of some parameters (e.g., estimating the slope of a linear trend). Although of major practical importance, estimates of complex functional relationships have received little laboratory or theoretical treatment. In particular, there has been no attempt to develop a scoring technique for measuring the excellence of such estimates.

In addition to the group I, event space E, and response space R, a Delphi exercise involves a process $G = G\,[I, E, R\,]$ which produces a group response G_j for each event E_j in the event space. Square brackets are used rather than parentheses in the expression for G to emphasize the fact that generally the group estimation process cannot be expressed as a simple functional relationship. The process may involve, for example, discussion among members of the group, other kinds of communication, iteration of judgments with complex selection rules on what is to be iterated, and so on.

One other piece of conceptual apparatus is needed, namely, the notion of score, or measure of performance. Development of scoring techniques has been slow in Delphi practice, probably because in most applied studies the requisite data for measuring performance either is unavailable, or would require waiting a decade or so. But in addition, the variety of subject matters, the diversity of motivations for applied studies, and the obscuring effect of the radical uncertainty associated with topics like long-range forecasting of social and technological events have inhibited the attempt to find precise measures of performance.

In the present paper, emphasis will be put on measures related to the accuracy of estimates. There is a large family of such measures, depending on the form of the estimate, and depending on the interests of the user of the estimate. For this essay, measures will be restricted to what might be called scientific criteria, i.e., criteria which do not include potential economic benefits to the user (or potential costs in terms of experts' fees, etc.) or potential benefits in facilitating group action.

For simple selections out of discrete event spaces a right/wrong measure is usually sufficient, for example, crediting the estimate with a 1 or 0 depending on whether it is correct or incorrect. However, as in the related area of performance testing

in psychology, the right/wrong measure is usually augmented by computing a score—total number right, or proportion right, or right-minus-wrong, etc.—over a set of estimates.

For simple selections out of continuous spaces (point estimates), a distance measure is commonly employed, for example, difference between the estimate and the true answer. However, if such measures are to be combined into a score over a set of estimates, some normalizing procedure must be employed to effect comparability among the responses. One normalizing procedure for always positive quantities such as dates, size of objects, probabilities, and the like, is the log error, defined as

$$\text{Error} = \log\left|\frac{R_i}{T}\right|,$$

where T is the true answer and R_i is the individual response. The vertical bars denote the absolute value (neglecting sign). Dividing by T equates proportional errors, and taking the logarithm uniformizes under- and over-estimates. Comparable scoring techniques have not been worked out for quantities with an inherent zero, i.e., quantities admitting both positive and negative answers. Such quantities are rare in applied exercises. Whether this is because that type of quantity is inessential to the subject matter or whether it is due to avoidance by practitioners is hard to say.

For probability estimates, some form of probabilistic scoring system appears to be the best measure available. The theory of probabilistic scoring systems is under rapid development. It is usually pursued within the ambit of subjective probability theories, where the primary property sought is a reward system which motivates the estimator to be honest, i.e., to report his "true" belief.

This requirement can be expressed as the condition that the expected score of the estimator should be a maximum when he reports his true belief. If $q = \{ q_j \}$ is the set of probabilities representing the actual beliefs of the estimator on event space $\{ E_j \}$, $R = \{ R_j \}$ is his set of reported probabilities, and $S_j(R)$ is the reward he receives if event E_j occurs, then the honesty condition can be written in the form:

$$\sum_j q_j S_j(q) \geqslant \sum_j q_j S_j(R). \qquad (1)$$

The expression on the left of the inequality is the individual's subjective expectation if he reports his actual belief; the expression on the right is his expectation if he reports something else.

Formula (1) defines a family of scoring (reward) systems often referred to as "reproducing scoring systems" to indicate that they motivate the estimator to reproduce his actual belief.

It is riot difficult to show that the theory of such scoring systems does not depend on the interpretation of q as subjective belief; it is equally meaningful if *q* is interpreted as the objective probability distribution *P* on *E*. With this interpretation the estimator is being rewarded for being as accurate as possible – his objective expectation is maximized when he reports the correct probability distribution.

This is not the place to elaborate on such scoring systems (see [1], [2], [3]). Although (1) leads to a family of reward functions, it is sufficient for the purposes of this essay to select one. The logarithmic scoring system

$$S_j(R) = A \log R_j + B \qquad (2)$$

has a number of desirable features. It is the only scoring system that depends solely on the estimate for the event which occurs. The expected score of the estimator is precisely the negative entropy, in the Shannon sense [4], of his forecast. It has the small practical difficulty that if the estimator is unfortunate enough to ascribe 0 probability to the alternative that occurs, his score is negatively infinite. This can usually be handled by a suitable truncation for very small probabilities.

Within this restricted framework, the Delphi design "problem" can be expressed as finding processes *G* which maximize the expected score of the group response. This is not a well-defined problem in this form, since the expectation may be dependent on the physical process being estimated, as well as on the group judgment process. There are two ways to skirt this issue. One is to attempt to find *G*'s which have some optimality property independent of the physical process. The other route is to assume that knowledge of the physical process can be replaced by knowledge about the estimators, i.e., knowledge concerning their estimation skill. The next section will deal with the second possibility.

There are two basic assumptions which underlie Delphi inquiries: (a) In situations of uncertainty (incomplete information or inadequate theories) expert judgment can be used as a surrogate for direct knowledge. I sometimes call this the "one head is better than none" rule. (b) in a wide variety of situations of uncertainty, a group judgment (amalgamating the judgments of a group of experts) is preferable to the judgment of a typical member of the group, the "n heads are better than one" rule.

The second assumption is more closely associated with Delphi than the first, which has more general application in decision analysis. These two assumptions do not, of course, exhaust all the factors that enter into the use of Delphi techniques. They do appear to be fundamental, however, and most of the remaining discussion in this paper will be concerned with one or the other of the two.

Individual Estimation

Using the expert as a surrogate for direct knowledge poses no problems as long as the expert can furnish a high-confidence estimate based on firm knowledge of his own. Issues arise when existing data or theories are insufficient to support a high-confidence estimate. Under these circumstances, for example, different experts are likely to give different answers to the same questions.

Extensive "everyday experience" and what limited experimental data exist on the subject strongly support the assumption that knowledgeable individuals can make useful estimates based on incomplete information. This general assumption, then, is hardly in doubt. What is in doubt is the degree of accuracy of specific estimates. What is needed is a theory of estimation that would enable the assignment of a figure of merit to individual estimates on the basis of readily available indices.

An interesting attempt to sidestep this desideratum is to devise methods of rewarding experts so that they will be motivated to follow certain rules of rational estimation. One approach to the theory of probabilistic scoring systems described in the introduction is based on this strategem [5].

The outlines of such a theory of estimation have been delineated in the literature of decision analysis; but it is difficult to disentangle from an attendant conceptualization of a prescriptive theory of decisionmaking, or as sometimes characterized, the theory of rational decisionmaking. In the following I will try to do some disentangling, but the subject is complex and is and ought may still intermingle more than one might wish.

In looking over the literature on decision analysis, there appear to be about six desirable features of estimation that have been identified. The number is not sharp, since there are overlaps between the notions and some semantic difficulties plague the classification. The six desiderata are honesty, accuracy, definiteness, realism, certainly, and freedom from bias.

Honesty is a clear enough notion. In most cases of estimation, the individual has a fairly distinct perception of his "actual belief," or put another way, he has a relatively clear perception whether his reported estimate matches his actual belief. This is not always the case. In situations with ambiguous contexts, such as the group-pressure situations created by Asch [6], some individuals appear to lose the distinction. The reason for wanting honest reports from estimators is also clear. Theoretically, any report, honest or not, is valuable if the user is aware of potential distortions and can adjust for them. But normally such information is lacking.

Accuracy is also a fairly straightforward notion, and is measured by the score in most cases. It becomes somewhat cloudy in the case of probability estimates for single events, where an individual can make a good score by chance. In this case, the average score over a sequence of events is more diagnostic. But the notion of accuracy then becomes mixed with the notion of realism. Given the meaningfulness of the term, the desirability of accuracy is clear.

Definiteness measures the degree of sharpness of the estimate. In the case of probabilities on discrete event spaces, it refers to the degree to which the probabilities approach 0 or 1 and can be measured by $\sum_{j=1}^{m} R_j^2$ In the case of probability distributions on continuous quantities, it can be measured by the variance or the dispersion. In the case of selections, the comparable notion is "refinement." For discrete event spaces, one report is a refinement of another if it is logically included in the second.

The reason for desiring definiteness is less clear than for accuracy or honesty. "Risk aversion" is a well-known phenomenon in economic theory, but "risk

preference" has also been postulated by some analysts [7]. In the case of discrete alternatives, the attractiveness of a report that ascribes a probability close to 1 to some alternative, and probability close to 0 to the others is intuitively "understandable." There is a general feeling that probabilistic: estimates close to 0 or 1 are both harder to make, and more excellent when made, than "wishy-washy" estimates in the neighborhood of 1/2. There is also the feeling that an individual who makes a prediction with a probability of .8 (and it turns out correct) knows more about the phenomenon being predicted than someone who predicts a similar event with probability .6.

All of this is a little difficult to pin down. In the experiments of Girshick, et al. [8], there was almost no correlation between a measure of definiteness and the accuracy of the estimates. Part of the problem here appears to be an overlap between the notion of definiteness and uncertainty, which is discussed below. At all events, there appears to be little doubt that definiteness is considered a virtue.

Realism refers to the extent that an individual's estimates are confirmed by events. It is thus closely related to accuracy. However, accuracy refers to a single estimate, whereas realism refers to a set of estimates generated by an individual. Other terms used for this notion are calibration [9], *precision* [10], *track record.*

Because the notion of realism is central to the first principle of Delphi stated in the introduction, namely, the substitution of expert judgment for direct knowledge, it warrants somewhat extensive discussion.

In the case of probability judgments, it is possible in theory to take a sequence of estimates from a single estimator, all with the same estimated probability, and count the number of times the estimate was confirmed. Presumably, if the estimator is using the notion of probability correctly, the relative frequency of successes in that sequence should be approximately equal to the estimated probability. Given enough data of this sort for a wide range of different estimates, it is possible in theory to generate a realism curve for each individual, as illustrated in Fig. 1.

In Fig. 1 the relative frequency with which an estimate of probability R_i is verified, $RF(C|R_i)$ *("C"* for "correct"), is plotted against the estimate. Realism can he defined as the degree to which the $RF(C|R_i)$ curve approximates the theoretically fully realistic curve, namely the dashed line in Fig. 1, where $RF(C|R_i) = R_i$. Figure 1 illustrates a typical realism curve where probabilities greater than z are "overestimated" and probabilities less than 1/2 are underestimated [11].

Various quantities can be used to measure the overall realism of an estimator. $\int_0^1 (RF(C|R_i) - R_i)^2 \; D(R_i)$ where D(R_i) is the distribution of the estimator's reports R_i – roughly the relative frequency with which he uses the various reports R_i – is a reasonable measure. However, for most applications of the concept, it is the realism curve itself which is of interest.

If such a curve were available for a given individual, it could be used directly to obtain the probability of a given event, based on his report. In particular, if the individual were fully realistic, the desired probability would be R_i. At first sight, it might appear that one individual, given his realism curve, is all that is needed to obtain a desired estimate, since the curve furnishes an "objective" translation of his reports

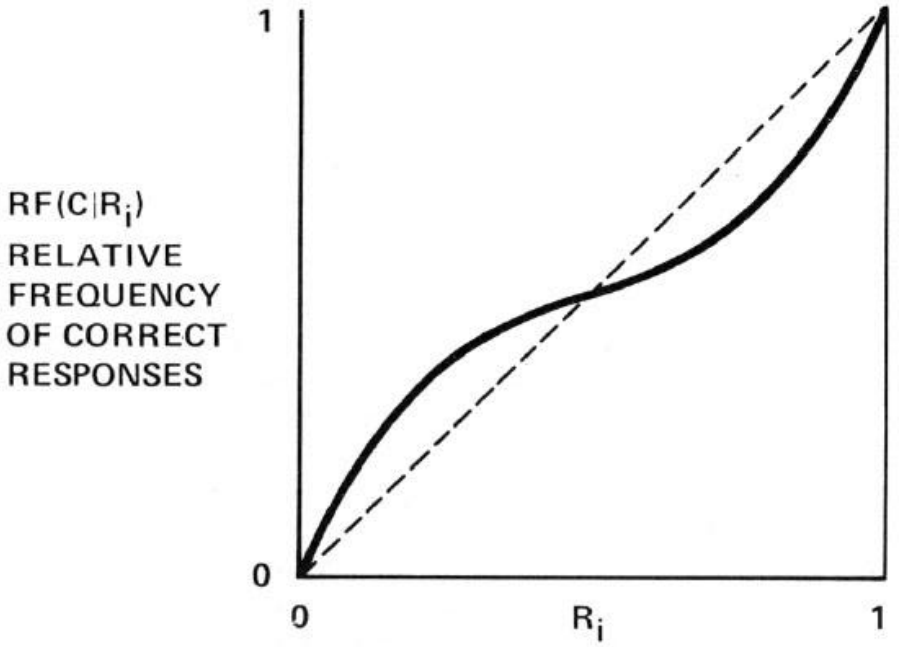

Fig. 1. Typical realism curve.

into probabilities. However, for any one specific estimate, the reports of several individuals typically differ, and in any case the realism curve is not, by itself, a measure of the expertness or knowledgeability of the individual. In particular, the frequency with which the individual reports relatively high probabilities has to be taken into account.

As a first approximation, the knowledgeability K_i. of individual i can be measured by

$$K_i = \int_0^1 S(R_i) D(R_i),$$

where $S(R_i)$ is the probabilistic score awarded to each report R_i and $D(R_i)$ is, as before, the distribution of the reports R_i.

It is easy to verify two properties of K_i: (a) K_i is heavily influenced by the degree of realism of the estimator. For a given distribution of estimates, $D(R_i)$, K_i is a maximum when the individual is fully realistic. (b) K_i is also influenced by the average definiteness of the estimator. The higher the definiteness (e.g., measured by $\int R_i^2 D(R_i)$), the higher the expected score.

Theoretically, one might pick the individual with the highest K rating and use him exclusively. There are two caveats against this procedure. On a given question, the

individual with the highest average *K* may not furnish the best response; and, more in the spirit of Delphi, if realism curves are available for a set of individuals, then it is sometimes feasible to derive a group report which will have a larger average score than the average score of any individual-in short, the K measure for the group can be higher than the K measure for any individual.

As far as the first principle-substitution of expert judgment for knowledge-is concerned, the question whether realism curves exist for each individual is a crucial one. Detailed realism curves have not been derived for the types of subject matter and the type of expert desired for applied studies. In fact, detailed track records for any type of subject matter are hard to come by. Basic questions are: Is there a stable realism curve for the individual for relevant subject matters? How general is the curve-i.e., is it applicable to a wide range of subject matters? How subject is the curve to training, to use of reward systems like the probabilistic score, to contextual effects such as the group pressure effect in the Asch experiments?

Certainty is a notion that is well known in the theory of economic decision-making. It has not played a role in the study of estimation to the same extent. In the case of economic decisionmaking, the distinction has been made between *risk* (situations that are probabilistic, but the probabilities are known) and *uncertainty* (situations where the probabilities are not known) [12]. Many analysts appear to believe that in the area of estimation this distinction breaks down-uncertainty is sufficiently coded by reported probabilities. However, the distinction appears to be just as applicable to estimation as to any other area where probabilities are relevant. Consider, for example, the situation of two coins, where an individual is asked to estimate the probability of *heads*. Coin A is a common kind of coin where the individual has flipped it several times. In this case, he might say that the probability of heads is ½ with a high degree of confidence. Coin B, let's say, is an exotic object with an unconventional shape, and the individual has not flipped it at all. In the case of coin B he might also estimate a probability of ½ for heads, but he would be highly uncertain whether that is the actual probability. Probability ½, then, cannot express the uncertainty attached to the estimate for the second coin.

A closer approximation to the notion of uncertainty can be obtained by considering a distribution on the probabilities. For example, the individual might estimate that the probability of the familiar coin has a tight distribution around ½, whereas the distribution for the unfamiliar coin is flat, as in Fig. 2. The independent variable is labeled q to indicate that it is the individual's belief, and not necessarily his report, which is being graphed.

The use of a higher-level distribution is only an approximation to the notion of uncertainty, since the distribution itself might be uncertain, or, in more familiar language, the distribution may be "unknown." The use of additional levels has been suggested, but for practical reasons seems highly unappealing.

The problem of representing uncertainty in precise terms is closely related to past attempts to translate lack of information into probabilities by means of principles such as the "law of insufficient reason," or the rule of equal ignorance. These have invariably lead to paradoxes [13].

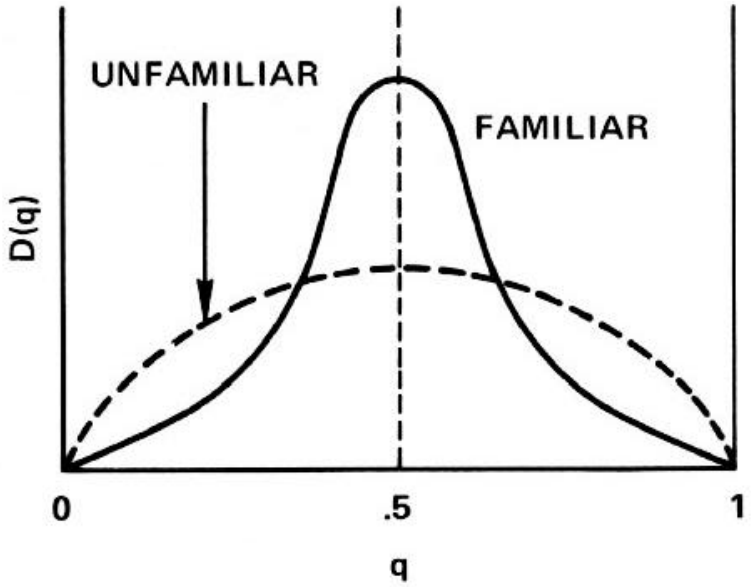

Fig. 2. Uncertainty represented as a higher-level distribution.

Using the idea of the dispersion of a second-level distribution as an approximate measure of uncertainty, there is some interaction between the notions of realism, definiteness, and certainty. It is not possible for a set of estimates to be simultaneously realistic, definite, and uncertain. Assuming that the individual will give as his first level report R_i the mean of his second-level distribution, then as R_i, approaches 1 or as R_i, approaches 0, the standard deviation of the distribution $D(q)$ approaches 0. Figure 3 illustrates this coupling for $R_i=0.9$. If the individual is realistic and estimates a probability of 0.9 for a given event, then the standard deviation of his higher-level distribution for that estimate must be small.

Unfortunately, the coupling applies only to the extremes of the 0 to 1 interval. At $q=$ ½, D can be about anything, and the estimator still be realistic. If the average probabilistic score for an estimate with a second-level distribution D is computed, the average score is influenced only by the mean, and otherwise is independent of D. Thus an average probabilistic score does not reflect uncertainty. It appears that something like the variance of the score will have to be included if certainty is to be reflected in a score.

At the present, the only "visible" index of certainty is the self-rating—i.e., a judgment by the individual of his competence or knowledgeability concerning the estimate. This has turned out to be a significant index for rating group estimates [14]; it is not so effective for individual estimates. Due to the lack of a theoretical definition of

the self-rating, it has not been possible to include it in a formal theory of aggregation.

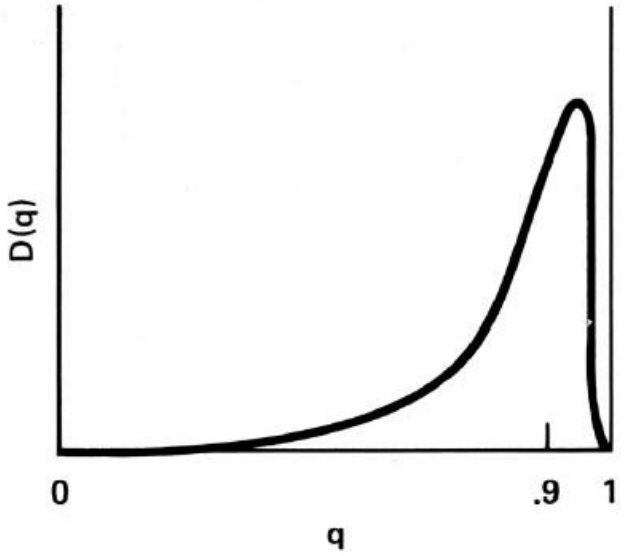

Fig. 3. Illustration of coupling between certainty and definiteness.

However, the self-rating has proved to be valuable for selecting more accurate subgroups [15].

Bias is a term that has many shades of meaning in statistics and probability. I am using the term to refer to the fact that there may be subclasses of events for which $RF(C \mid R_i)$ may be quite different from the average relative frequency expressed by the realism curve. Of course, for this to be of interest, the subclasses involved must be identifiable by some means other than the relative frequency. It is always possible after the fact to select a subset of events for which an individual has estimated the probability R_i which has any $RF(C \mid R_i)$.

In the theory of test construction, e.g., for achievement tests or intelligence tests, it is common to assume an underlying scale of difficulty for the questions, where difficulty is defined as the probability that a random member of the target population can answer the question correctly [16]. This probability will range from 1 for very easy questions to 0 for very hard questions, as illustrated by the solid curve in Fig. 4. From the standpoint of the present discussion, the significant fact is that when a class of questions is identified as belonging to the very difficult group in a sample of the population, that property carries over to other members of the population—in short the property of being very difficult is relatively well defined.

At some point in the scale of difficulty, labeled *d* in Fig. 4, a typical member of the population could increase his score by abandoning the attempt to "answer" the question and simply flipping a coin (assuming that it is a true/false or yes/no type of question). Put another way, from point d on, the individual becomes a counterpredictor—you would be better off to disbelieve his answers.

Contrasted with this notion of difficulty is the notion that underlies theories of subjective probability that, as the individual's amount of information or skill declines, the probability of a correct estimate declines to 50 percent as illustrated by the dashed curve in Fig. 4. Ironically, it is the probabilistic notion that influences most scoring schemes, which assume that the testee can achieve 50 percent correct by "guessing," and hence the score is computed by subtracting the number of wrong answers from the number right. By definition, for the more difficult items, the testee cannot score 50 percent by "guessing" unless that means literally tossing a coin and not trusting his "best guess."

If it turns out that "difficult" questions in the applied area have this property, even for experts, then the first principle does not hold for this class. Although there are no good data on this subject, there does not appear to be a good reason why what holds for achievement and intelligence tests should not also hold for "real life" estimates. Almost by definition, the area of most interest in applications is the area of difficult questions. If so, assuming that the set of counterpredictive questions can be identified before the fact, then a good fair coin would be better than an expert.' It is common in experimental design to use randomization techniques to rule out potential biases. There *is* no logical reason why randomization should not be equally potent in ruling out bias in the case of estimation.

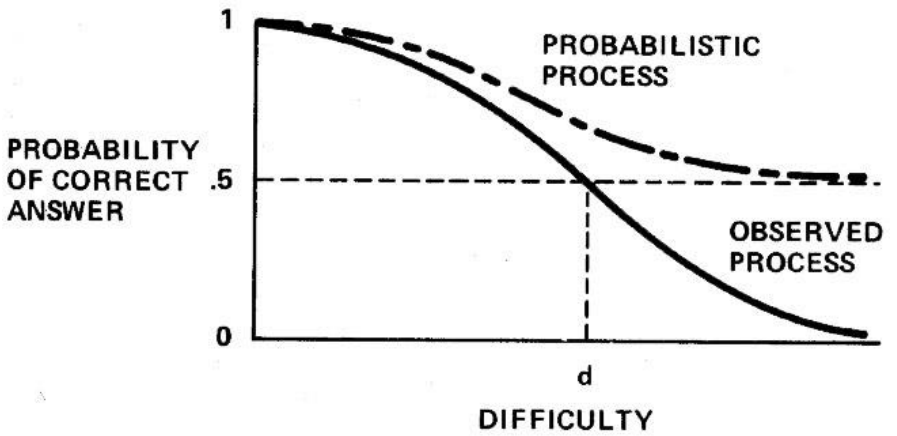

Fig. 4. Scale of difficulty in test construction.

The four notions, honesty, accuracy, definiteness, and precision, are all tied together by probabilistic scoring systems. In fact, a reproducing scoring system rewards the estimator for all four. As pointed out in the introduction, condition (1) defines the same family of scoring systems whether q is interpreted as subjective belief, or as objective probability. Thus, the scoring system rewards the estimator for both honesty and accuracy. In addition, the condition leads to the result that

$\Sigma_j q_j S_j(q)$ is convex in q. This convex function of q can be considered as a measure of the dispersion of q; and in fact, three of the better-known scoring systems define three of the better-known measures of dispersion. Thus, if S_j is the quadratic scoring system

$$S_j(R) = 2R_j - \sum_j R_j^2,$$

then $\Sigma_j q_j S_j(q) = \Sigma_j q_j^2$ which is a measure of variance. If S_j is the spherical scoring system:

$$S_j(R) = \frac{R_j}{\sqrt{\sum_j R_j^2}},$$

then $\Sigma_j q_j S_j(q) = \sqrt{\Sigma_j q_j^2}$, a measure similar to the standard deviation. Finally, for the logarithmic scoring system $\Sigma_j q_j S_j(q) = \Sigma_j q_j \ln q_j$ which is the negative of the Shannon entropy, another measure of definiteness.

Realism enters in a more diffuse fashion. In general, the probabilistic score for a single event is not very diagnostic, since the individual may have obtained a high (or low) score by chance. Thus, as for most scoring systems, an average (or total) score over a large set of questions is the usual basis for evaluation. But over a large set of questions, the average score is determined by the realism curve of the individual" in conjunction with the relative frequency with, which he makes reports *of* a *given* probability. In general, if the estimator is not realistic, he will lose

$$\int_0^1 (RF(C|R_i)S(RF(C|R_i)) - RF(C|R_i)S(R_i))D(R_i).$$

As pointed out above, the probabilistic score does not include a penalty for uncertainty, nor does it include a penalty for bias, except where bias shows up *in* the realism curve. The latter case is simply the one where, for whatever reason, the individual is faced with a stream of questions in which the number of questions biased in *a* given direction is greater than the number biased in the opposite direction.

To sum up this rather lengthy section: The postulate that, in situations of uncertainty, it is feasible to substitute expert judgment for direct knowledge is

grounded in a number of empirical hypotheses concerning the estimation process. These assumptions are, primarily, that experts are approximately realistic in the sense defined above, that the realism curve is stable over a relatively wide range of questions (freedom from bias), and that knowledgeability is a stable property of the expert. At the moment, these are hypotheses, not well-demonstrated generalizations.

Theoretical Approaches to Aggregation

Assuming that, for a given set of questions, we can accept the postulate that expert judgment is the "best information obtainable," there remains the question how the judgments of a group of experts should be amalgamated. In the present section, three approaches to this issue are discussed. The discussion is limited to elementary forms of aggregation, where the theory consists of a mathematical rule for deriving a group response from a set of individual responses; thus, an *elementary* group estimation process can be defined as a function, $G = G(E, I, R)$.

Theory of Errors

This approach interprets the set of judgments of a group of experts as being similar to the set of readings taken with an instrument subject to random error. It seems most appropriate when applied to point estimates of a continuous quantity, but formally at least, can be applied to any type of estimate. In analogy with the theory of errors for physical measurements, a statistical measure of central tendency is considered to be the best estimate of the quantity. Some measure of dispersion is taken to represent a confidence interval about the central value.

Relevant aspects of the individual estimation process such as skill or amount of information of the expert, are interpreted as features of the "theory of the instrument."

This point of view appears to be most popular in the Soviet Union [17]; however, a rough though unexpressed version of this approach underlies much of the statistical analysis accompanying many applied Delphi studies. To my knowledge, this approach has not been developed in a coherent theory, but rather, has been employed as an informal "interpretation"— i.e., as a useful analogy.

The theory-of-errors approach has the advantages of simplicity, and similarity with well-known procedures in physical measurement theory. Much of the empirical data which have been collected with almanac and short-range prediction studies is compatible with the analogy. Thus, the distribution of estimates tends to follow a common form, namely the lognormal [18]. If the random errors postulated in the analogy are assumed to combine multiplicatively (rather than additively as in the more common Gaussian theory), then a lognormal distribution would be expected.

The geometric mean of the responses is more accurate than the average response; or more precisely, the error of the geometric mean is smaller than the average error. Since the median is equal to the geometric mean for a lognormal distribution [19], the median is a reasonable surrogate, and has been the most widely used statistic in applied studies for the representative group response.

The error of the median is, on the average, a linear function of the standard deviation [20], which would be predicted by the theory of errors. The large bias observed experimentally (bias= error/standard deviation) is on the average a constant, which again would be compatible with the assumption that experts perform like biased instruments.

Although the analogy looks fairly good, there are several open questions that prevent the approach from being a well-defined theory. There does not exist at present a "theory of the instrument" which accounts for either the observed degree of accuracy of individual estimates or for the large biases observed in experimental data. Perhaps more serious, there is no theory of errors which accounts for the presumed multiplicative combination of errors-especially since the "errors" are exemplified by judgments from different respondents.

Despite this lack of firm theoretical underpinnings, the theory-of-errors approach appears to fit the accumulated data for point estimates more fully than any other approach.

In addition, the measures of central tendency "recommended by" the theory of errors have the desirable feature that the advantage of the group response over the individual response can be demonstrated irrespective of the nature of the physical process being estimated. So far as I know, this is the only theoretical approach that has this property.

To make the demonstration useful in later sections, a somewhat more sophisticated version of the theory will be dealt with than is necessary just to display the "group effect."

Consider a set of individual estimates R_{ij} on an event space E_j, where the R_{ij} are probabilities, i.e., $\Sigma_j R_{ij} = 1.$. We assume there is a physical process that determines objective probabilities $P = \{ P_j \}$ for the event space, but P is unknown. Consider a group process G which takes the geometric mean of the individual estimates as the best estimate of the probability for each event. However, the geometric means will not be a probability, and must be normalized. This is accomplished by setting

$$G_j = \frac{\left(\prod_{i=1}^{n} R_{ij} \right)^{\frac{1}{n}}}{\sum_{j=1}^{m} \left(\prod_{i=1}^{n} R_{ij} \right)^{\frac{1}{n}}}. \qquad (2)$$

We can now ask how the expected probabilistic score of the group will compare with the average expected score of the individual members of the group. It is convenient to use the abbreviation C for the reciprocal of the normalizing term

$$C = \frac{1}{\sum_{j=1}^{m} \left(\prod_{i=1}^{n} R_{ij} \right)^{\frac{1}{n}}}.$$

Using the logarithmic scoring system and setting the constants A =1, B =0, we have:

$$S_j(G) = \log\left(C \left(\prod_{i=1}^{n} R_{ij} \right)^{\frac{1}{n}} \right) \tag{3}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \log R_{ij} + \log C. \tag{4}$$

Taking the expected score,

$$\sum_{j=1}^{m} P_j S_j(G) = \sum_{j=1}^{m} P_j \frac{1}{n} \sum_{i=1}^{n} \log R_{ij} + \log C, \tag{5}$$

and rearranging terms, where $\bar{S}(G)$ denotes the expected score of the group, i.e., $\bar{S}(G) = \Sigma_{j=1}^{m} P_j S_j(G)$

$$\bar{S}(G) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} P_j \log R_{ij} + \log C, \tag{6}$$

log C appears outside the summarion, because, as a constant, $\Sigma_{j=1}^{m} P_j \log C = \log C$. The expression $\Sigma_{j=1}^{m} P_j \log R_{ij}$ is just the expected score $\bar{S}(R_i)$ of individual i, and the expression on the right of (6) excluding log C is the *average* individual expected score, which we can abbreviate as $\hat{S}(R)$. Thus

$$\bar{S}(G) = \hat{S}(R) + \log C. \tag{7}$$

Since C is greater than 1, log C is positive, and the expected group score is greater than the average expected individual score by the amount log C. C depends only on the individual responses R_{ij} and not on the specific events E or the objective probabilities P.

Formula (7) exemplifies a large variety of similar results that can be obtained by using different statistics as the aggregation rule and different scoring rules.[1]

Probabilistic Approach

Theoretically, joint realism curves similar to the individual realism curve of Fig. 1 can be generated, given enough data. In this case, the relative frequency *RF(C|R) of* correct estimates would be tabulated for the joint space of responses *R* for a group. Such a joint realism curve would be an empirical aggregation procedure. *RF (C|R)* would define the group probability judgment as a function of *R*..

Although possible in theory (keeping in mind all the caveats that were raised with respect to individual realism curves), in practice generating joint realism curves for even a small group would be an enormous enterprise. It is conceivable that a small group of meteorologists, predicting the probability of rain for a given locality many thousands of times, might cover a wide enough region of the *R* space to furnish stable statistics. However, for the vast majority of types of question where group estimation is desired, individual realism curves are difficult to come by; group realism curves appear to be out of the question for the present.

One possible simplification at this point could be made if general rules concerning the interdependence of individual estimates on various types of estimation tasks could be ascertained. In such a case, the joint realism curves could be calculated from individual realism' curves. Although very iffy at this point, it is conceivable that a much smaller body of data could enable the testing. of various hypotheses concerning dependence. In any case, by developing the mathematical relationships involved, it is possible to pursue some theoretical comparisons of probabilistic aggregation with other types of aggregation.

In the following, the convention will be used that whenever the name of a set of events occurs in a probability expression, it denotes the assertion of the joint occurrence of the members of the set. For example, if X is a set of events, $X = \{X_i\}$, then $P(X) = P(X_1 * X_2 \ldots X_n)$, where the period indicates "and." In addition, to reduce the number of subscripts, when a particular event out of a set of events is referred to, the capital letter of the index of that event will be used to refer to the occurrence of the event. Thus $P(X_j)$ will be written $P(J)$.

The degree of dependence among a set of events X is measured by the departure of the joint probability of the set from the product of the separate probabilities of the events. Letting D_x denote the degree of dependence within the set X, we have the definition

[1] Brown [21] derives a similar result for continuous distributions, the quadratic scoring system, and the mean as the group aggregation function.

$$D_X = \frac{P(X)}{\prod_{i=1}^{n} P(X_i)}. \tag{8}$$

This notion is usually introduced by taking into account dependence among subsets of X as well as the more global notion defined by (8). However, for generating a probabilistic aggregation function, interactions among subsets can be ignored, proving we maintain a set fixed throughout any given computation.

A useful extension of the notion of dependence is that of dependence with respect to a particular event, say $E_j = J$.

$$D_X^J = \frac{P(X|J)}{\prod_{i=1}^{n} P(X_i|J)}. \tag{9}$$

From the rule of the product, we have

$$D_X^J = \frac{P(J \cdot X) P(J)^{n}}{P(J) \prod_{i=1}^{n} (J \cdot X_i)}. \tag{10}$$

The probability we want to compute is $P(J \mid R)$; that is, we want to know the probability of an event E_j, given that the group reports R. Again, from the rule of the product, we have

$$P(J|R) = \frac{P(R \cdot J)}{P(R)}.$$

Substituting R for X in (10) and multiplying the top and bottom of the right-hand side by $P(R)/\prod_{i=1}^{m} P(R_i)$, and rearranging, gives

$$P(J|R) = \frac{D_R^J \prod_{i=1}^{n} P(J|R_i)}{D_R U(J)^{n-1}}. \tag{11}$$

Formula (11) presents the computation of the joint probability in terms of the individual reports, the dependency terms, and the "a priori" probability $U(J)$. The $P(J \mid R_i)$ can be derived from individual realism curves. In case the estimators are all fully

realistic, then $P(J \mid R_i) = R_i$. $U(J)$ is the probability of the event J based on whatever information is available without knowing R.[2]

The ratio D_R^J / D_R measures the extent to which the event J influences the dependence among the estimates. If the estimates are independent "a priori," $D_R = 1$. However, the fact that estimators do not interact (anonymity) or make separate estimates, does not guarantee that their estimates are independent.

They could have read the same book the day before. The event related dependence D_R^J is even more difficult to derive from readily available information concerning the group.

If there is reason' to believe that a particular. group is completely independent in their estimates, and in addition each mdmber is completely realistic, (11) reduces to

$$P(J|R) = \frac{\prod_{i=1}^{n} R_i}{U(J)^{n-1}}. \tag{12}$$

The simplicity of (12) is rather misleading; it depends on several strong assumptions. (11) on the other hand, is exact, but contains terms which are difficult to evaluate.

An exact expression for $P(J \mid R)$ can be obtained which does not involve D_R by, noting that

$$P(J|R) = \frac{P(J|R)}{P(J|R) + P(\bar{J}|R)}.$$

Substituting for $P(J \mid R)$ on the right-hand side from (11) and the corresponding expression for $P(\bar{J}|R)$ ($\bar{J}$ denotes the complement of J or "not-J") and dividing top and bottom by $D_J^R / U(J)^{n-1} D_R$ we obtain

$$P(J|R) = \frac{\prod_{i=1}^{n} P(J|R_i)}{\prod_{i=1}^{n} P(J|R_i) + D \prod_{i=1}^{n} (1 - P(J|R_i))}, \tag{13}$$

[2] AII of the formulations in this subsection are presumed to be appropriate for some context of *information.* This context could be included in the formalism, e.g., as an additional term in the reference class for all relative probabilities, or as a reference class for "absolute" probabilities. For example, if the context is labeled W, $U(J)$ would be written $P(J \mid W)$, $P(J \mid R)$ would be written $P(J \mid R*W)$. However, since W would be constant throughout, and ubiquitous in each probability expression, it is omitted for notational simplicity.

where

$$D = \frac{D_R^{\bar{J}}}{D_R^{J}} \left(\frac{U(J)}{U(\bar{J})} \right)^{n-1}$$

If the estimators are all fully realistic and fully independent, and the a priori probability= ½, (13) reduce to

$$P(J|R) = \frac{\prod_{i=1}^{n} R_i}{\prod_{i=1}^{n} R_i + \prod_{i=1}^{n} (1-R_i)}. \tag{14}$$

To complete this set of estimation formulae, if there are several alternatives in E, and it is desired to compute the group estimate for each alternative from the individual estimates for each alternative, (13) generalize to

$$P(E_k|R) = \frac{\prod_{i=1}^{n} P(E_k|R_i)}{\sum_{j=1}^{m} D_{jk} \prod_{i=1}^{n} P(E_j|R_i)}, \tag{15}$$

where

$$D_{jk} = \frac{D_R^{E_j}}{D_R^{E_k}} \left(\frac{U(E_k)}{U(E_j)} \right)^{n-1}.$$

(14) is similar to a formula that can be derived using the theorem of Bayes [22]. Perhaps the major difference is that (14) makes the "working" set of estimates the $P(E_j \mid R_i)$ which can be obtained directly from realism curves, whereas the corresponding formula derived from the theorem of Bayes involves as working estimates $P(R_i \mid E_j)$ which are not directly obtainable from realism curves. Of course, in the strict sense,, the two formulae have to be equivalent, and the $P(R_i \mid E_j)$ are contained implicitly in the dependency terms. Without some technique for estimating the dependency terms separately from the estimates themselves, not much is gained by computing the group estimate with (14).

Historically, the "a priori" probabilities $U(J)$ have posed a number of conceptual and data problems to the extent that several analysts, e.g., R. A. Fisher [23], have

preferred to eliminate them entirely and work only with the likelihood ratios-in the case of (14), the ratios

$$\frac{\prod_{i=1}^{n} R(E_j|R_i)}{\prod_{i=1}^{n} P(E_k|R_i)}.$$

This approach appears to be less defensible in the present case, where the a priori probabilities enter in a strong fashion, namely with the n -1 power.

For a rather restricted set of situations, a priori probabilities are fairly well defined, and data exist for specifying them. A good example is the case of weather forecasting, where climatological data form a good base for a priori probabilities. Similar' data exist for trend forecasting, where simple extrapolation models are a reasonable source for a priori probabilities. However, in many situations where expert judgment is desired, whatever prior information exists is in a miscellaneous form unsuited for computing probabilities. In fact, it is in part for precisely this reason that experts are needed to "integrate" the miscellaneous information:

Some additional light can be thrown on the' role of a priori probabilities as well as the dependency terms by looking at the expected probabilistic score. In the case of the theory-of-errors approach, it was possible to derive the result that, independent of the objective probability distribution *P*, the expected probabilistic score of the group estimate is higher than the average expected score of individual members,of the group. This result is not generally true for probabilistic aggregation.

Since probabilistic aggregation depends upon knowing the a priori probabilities, a useful way to proceed is to define a *net* score obtained by subtracting the score that would be obtained by simply announcing the a priori probability. Letting $S^*(G)$ denote the expected net score of the group and $S^*(R_i)$ the expected net score of individual i, and $S(E)$ the score that would be obtained if $\{U(E_j)\}$ were the report, $S^*(G)=S(G)-S(E)$ and $S^*(R_i)=S(R_i)-S(E)$. The net score measures the extent to which the group estimate is better (or worse) than the a priori estimate. This appears to be a reasonable formulation, since presumably the group has' added nothing if its score is no better (or is worse) than what could be obtained without it.

Many formulations of probabilistic scores include a similar consideration when they are "normalized." This is equivalent to subtracting a score for the case of equally distributed probabilities over the alternatives. Thus the score for an individual is normalized by setting $S^*(R_i) = S(R_i) - S(Q)$ where $Q_j = 1/m$ and m is the number of alternatives. In effect this is assuming that the a priori probabilities are' equal.

Computing the expected group net score from (11) we have

$$\sum_{j=1}^{m} P_j G_j - S(E) = \sum_{j=1}^{m} P_j \ln\left(\frac{D_R^J \prod_{i=1}^{n} P(J|R_i)}{(J)^{n-1} D_R} \right) - S(E) \tag{16}$$

$$= -(n-1) \sum_{j=1}^{m} P_j \ln U(J) + \sum_{j=1}^{m} P_j \sum_{i=1}^{n} \ln P(J|R_i)$$

$$+ \sum_{j=1}^{m} P_j \ln D_R^J - \ln D_R - S(E) \tag{17}$$

$$= -nS(E) + n\hat{S}(R) + \sum_{j=1}^{m} P_j \ln D_R^J - \ln D_R, \tag{18}$$

whence $S^*(G) = nS^*(R)$ + Expectation of dependency terms.

If the average net score of the individual members is positive (i.e., the average member of the group does better than the a priori estimate), then the group score will be n times as good, providing the dependency terms are small

or positive. On the other hand, if the average net score of the individual members is negative, then the group will be n times as bad, still assuming the dependency terms small. Since the logarithm of D_R will be negative if $D_R < 1$, (18) shows that the most favorable situation is not independence where $D_R = 1$, In $D_R = 0$, but rather, the case of negative dependence, i.e., the case where it is less likely that the group will respond with R than would be expected from their independent frequencies of use of R_i.

The role of the event-related dependency term $\Sigma_{j=1}^{n} P_j \ln D_R^J$ is somewhat more complex. In general, it is desirable that D_R^J be greater than one for those alternatives where the objective probability P is high. This favorable condition would be expected if the individuals are skilled estimators, but cannot be guaranteed on logical grounds alone.

One of the more significant features of the probabilistic approach is that under favorable conditions the group response can be more accurate than any member of the group. For example, if the experts are fully realistic, agree completely on a given estimate, are independent, and finally, if it is assumed that the a priori probabilities are equal (the classic case of complete prior ignorance), then formula (14) becomes

$$P(J|R) = \frac{p^n}{p^n + (1-p)^n}, \qquad (19)$$

where p is the common estimate, and n is the number of members of the group. If $p>.5$, then $P(J|R)$ rapidly approaches 1 as n increases. For example, if $p=2/3$ and n is 5, then $P(J|R)=32/33$. If the theory-of-errors approach were being employed, the group estimate would be 2/3 for any size group.

In this respect, it seems fair to label the probabilistic approach "risky" as compared with the theory-of-errors approach. Under favorable conditions the former can produce group estimates that are much more accurate than the individual members of the group; under less favorable conditions, it can produce answers which are much worse than any member of the group.

Axiomatic Approach

A somewhat different way to develop a theory of group estimation is to postulate a set of desired characteristics for an aggregation method and determine the process or family of processes delimited by the postulates. This approach has not been exploited up to now in Delphi research. "I he major reason has been the large number of nonformal procedures associated with an applied Delphi exercise--formulation of a questionnaire, selection of a panel of experts, "interpretation of results," and the like. However, if the aggregation process is defined formally as in the two preceding subsections, where questionnaire design is interpreted as defining the event space E, and panel selection; is reduced to, defining the response space R, then the axiomatic approach becomes feasible.

Considering the group estimation process as -a function $G=G(E,J,R)$, various properties of this function appear "reasonable" at first glance. Some of the more evident of these are:

(A) *Unanimity.* If the group is *in* complete agreement, then the group estimate is equal to the common individual estimate; i.e., if $R_{ij} = R_{kj}$ for all i and k, then $G(R)=R$.

(B) *Monotony.* If R and R' are such that $R_{ij} \geq R'_{ij}$ for all i, then $G_j(R) \geq G_j(R')$. *If R* and G are defined as real numbers then they fulfill the usual ordering axioms, and condition *B* implies condition *A.*

(C) *Nonconventionality. G is* not independent of the individual estimates; i.e., $G(R) \geq G(S)$ for every possible R and S.

(D) *Responsiveness. G* is responsive to each of the individual estimates; i.e., $G(R) \geq G'(T)$, where T *is* a proper subvector of *R.*

(E) *Preservation of Probability Rules.* If *G is* an aggregation function which maps a set of individual probability estimates onto a probability, then G preserves the rules of probability. For example, if $T_{ij} = R_{ij}\, S_{ij}$, for all i and j (as would be the case if R_{ij} *is* the estimated probability of E_j and S_{ij} is the estimated relative probability of an event E'_j given that E_j occurs) then

$$G(T_j) = G(R_j)G(S_j).$$

This set of conditions will be displayed more fully below.

All of these conditions have a fairly strong intuitive appeal. However, intuition appears to be a poor guide here. The first four postulates are fulfilled by any of the usual averaging techniques. But *A*, which is perhaps the most apparently reasonable of them all, is not fulfilled by the probabilistic aggregation techniques discussed in the previous subsection. It was pointed out there that one of the more intriguing possibilities with probabilistic aggregation is that the group estimate may be higher (or lower, depending on the interaction terms) than any individual estimate.

It can be shown that there is no function that fulfills all five of the postulates; in fact, there is *no* function that fulfills *D* and *E*. The proof of this impossibility theorem is given elsewhere [24]; it will *only* be sketched here.

Three basic properties of probabilities are *(a)* normalization, if *p* is a probability, 0 ? *p* ? 1; *(b)* complementation, $P(J)+P(\bar{J})=1$; and *(c)* multiplicative conjunction, i.e., $P(J_1*J_2)=P(J_1)P(J_2/J_1)$. The last is sometimes taken as a postulate, sometimes is derived from other assumptions.

If the individual members of a group are consistent, their probability judgments will fulfill these three conditions. It would appear reasonable to require that a group estimate also fulfill the conditions, consistently with the individual judgments. In addition, condition *D*, above, appears reasonable. This leads to the four postulates:

P1. $0 \leqslant G(R) \leqslant 1.$
P2. $G(1-R)=1-G(R).$
P3. $G(R\cdot S)=G(R)G(S).$
P4. $G(R)\neq G'(S)$, where S is a subvector of R (condition D).

Here, $R\cdot S$ is the inner product of the two vectors R and S, i.e.,

$$R\cdot S=(R_1S_1, R_2S_2, \ldots, R_nS_n).$$

Pl-P3 have the consequence that *G* is both multiplicative and additive. The multiplicative property comes directly from P3, and the additive property—i.e., $P(R+S)=P(R)+P(S)$ – is derived by using the other postulates. For functions of a single variable, there is only one which is both multiplicative and additive, namely the identity function $f(x)=x$. There is no corresponding identity function for functions of several variables except the degenerative function, $G(R) = G'(R_i) = R_i$, which violates P4.

This result may seem a little upsetting at first glance. It states that probability estimates arrived at by aggregating a set of individual probability estimates cannot be manipulated as if they were direct estimates of a probability. However, there are many ways to react to an impossibility theorem. One is panic. There is the story that the logician Frege died of a heart attack shortly after he was notified by Bertrand Russell of the antinomy of the class of all classes that do not contain themselves. There was some such reaction after the more recent discovery of an impossibility theorem in the area of group preferences by Kenneth Arrow [25]. However, a quite different, and more pragmatic reaction is represented by the final disposition of the case of 0. In the 17th century, there was long controversy on the issue whether 0 could be treated as a number. Strictly speaking there is an impossibility theorem to the effect that 0 cannot be a number. As everyone knows, division by 0 can lead to contradictions. The resolution was a calm admonition, "Treat 0 as a number, but don't divide by it."

In this spirit, formulation of group probability estimates has many desirable properties. It would be a pity to forbid them because of a mere impossibility theorem. Rather, the reasonable attitude would appear to be to use group probability estimates, but at the same time not to perform manipulations with the group aggregation function which can lead to inconsistencies.

Coda

The preceding has taken a rather narrow look at some of the basic aspects of group estimation. Many significant features, such as interaction via discussion or formal feedback, the role of additional information "fed-in" to the group, the differences between open-ended and prescribed questions, and the like, have not been considered. In addition, the role of a Delphi exercise within a broader decisionmaking process has not been assessed. What has been attempted, albeit not quite with the full neatness of a well-rounded formal theory, is the analysis of some of the basic building blocks of group estimation.

To summarize briefly: The outlines of a theory of estimation have been sketched, based on an objective definition of estimation skill-the realism curve or track record of an expert. Several approaches to methods of aggregation of individual reports into a group report have been discussed. At the moment, insufficient empirical data exist to answer several crucial questions concerning both individual and group estimation.[11] For individual estimation, the question is open whether the realism curve is well defined and sufficiently stable so that it can be used to generate probabilities. For groups, the degree of dependency of expert estimates, and the efficacy of various techniques such as anonymity and random selection of experts in reducing dependency have not been studied.

By and large it appears that two broad attitudes can be taken toward the aggregation process. One attitude, which can be labeled conservative, assumes that expert judgment is relatively erratic and plagued with random error. Under this assumption, the theory-of-errors approach looks most appealing. At least, it offers the comfort of the theorem that the error of the group will be less than the average error of the individuals. The other attitude is that experts can be calibrated and, via training and

computational assists, can attain a reasonable degree of realism. In this case it would be worthwhile to look for ways to obtain a priori probabilities and estimate the degree of dependency so that the more powerful probabilistic aggregation techniques can be used.

At the moment I am inclined to -take the conservative attitude because of the gaping holes in our knowledge of the estimation process. On the other hand, the desirability of filling these gaps with extensive empirical investigations seems evident.

References

1. John McCarthy, "Measures of the Value of Information," Proc. *Nat. Acad. of Sci.* 42 (September 15, 1956), pp. 654-55.
2. Thomas Brown, "Probabilistic Forecasts and Reproducing Scoring Systems," The Rand Corporation, RM-6299-ARPA, July 1970.
3. L. J. Savage, "Elicitation of Personal Probabilities and Expectations," J. Amer. Slat. Assoc. 66 (December 1971), pp. 783-801.
4. E. E. Shannon and W. Weaver, *The Mathematical Theory of Communication,* University of Illinois Press, Urbana, 1949.
5. Savage, op. cit.
6. S. E. Asch, "Effects of Group Pressure upon the Modification and Distortion of Judgments," in E. E. Maccoby, T. M. Newcomb, and E. L. Hartley (eds.), *Readings in Social Psychology,* Henry Holt, New York, 1958, pp. 174-83.
7. C. H. Coombs, "A Review of the Mathematical Psychology of Risk," presented at the Conference on Subjective Optimality, University of Michigan, Ann Arbor, August 1972.
8. M. Girshick, A. Kaplan, and A. Skogstad, "The Prediction of Social and Technological Events," Public Opinion Quarterly, Spring 1950, pp. 93-110.
9. G. A. S. Stael von Holstein, "Assessment and Evaluation of Subjective Probability Distributions," Economic Research Institute, Stockholm School of Economics, 1970.
10. Girshick, et al., op. cit.
11. W. Edwards, "The Theory of Decision Making," *Psychol. Bulletin* 5 (1954), pp. 380-417.
12. F. Knight, *Risk, Uncertainty and Profit,* Houghton Mifflin, Boston, 1921.
13. Hans Reichenbach, *The Theory of Probability,* University of California Press, Berkeley, 1949, Section 68.
14. N. Dalkey, "Experimental Study of Group Opinion," *Futures 1* (September 1969), pp. 408-26.
15. N. Dalkey, B. Brown, and S. Cochran, "The Use of Self-Ratings to Improve Group Estimates," *Technological Forecasting 1* (1970) pp. 283-92.
16. J. P. Guilford, *Psychometric Methods,* McGraw-Hill, New York, 1936, pp. 426ff.
17. N. Moiseev, "The Present State of Futures Research in the Soviet Union," in *Trends* in *Mathematical Modeling,* Nigel Hawkes (ed.), Springer-Verlag, Berlin, 1973.
18. N., Dalkey, "Experimental Study of Group Opinion," op. cit.
19. J. Aitchison and J. A. C. Brown, *The Lognormal Distribution,* University of Cambridge Press, Cambridge, Eng., 1957.
20. N. Dalkey, "Experimental Study of Group Opinion," op. cit.
21. T. ; Brown, "An Experiment in Probabilistic Forecasting," The Rand Corporation, .R-944-ARPA, March 1973.
22. Peter A. Morris, *Bayesian Expert Resolution,* doctoral dissertation, Stanford University, Stanford, California, 1971.

23. R. A. Fisher, "On the Mathematical Foundations of Theoretical Statistics," Philos. *Trans. Roy.* Soc., London, Series A, Vol. 222, 1922.
24. N. Dalkey, "An Impossibility Theorem for Group Probability Functions," The Rand Corporation P-4862; June 1972.
25. K. J. Arrow, *Social Choice and Individual Values,* John Wiley and Sons, New York, 1951.