The Utilization of Expert Opinion in Decision-Making

This is a review of the state-of-the-art of rational and primarily quantitative approaches to the utilization of expert opinion in the inexact sciences, among which the engineering sciences must be included. Of primary concern is the decision-maker's utilization of one or more experts, and the criteria and scoring rules used to evaluate their performance. Also reviewed are certain game-theoretic strategies used by experts to maximize their own goal satisfaction. With this review, it is hoped that chemical engineers, when acting as decision-makers, will be able to utilize expert opinion more effectively and will better understand the function of experts when it is necessary for them to adopt such roles themselves.

ALBERT J. KLEE

Solid Waste Research Laboratory
National Environmental Research Center
Environmental Protection Agency
Cincinnati, Ohio 45268

SCOPE

In many decision-making situations, the use of experts plays a vital role in the process. The reasons are not hard to find. The decision-maker, for example, may himself have very little knowledge of the pertinent subject matter; alternatively, he may lack the analytical capacity to evaluate the problems before him. In the inexact sciences (for example, sociology, psychology, economics, systems analysis, meteorology, engineering, and the environmental sciences in general), the vagueness of certain background knowledge, which nonetheless may be significant or even predominantly important, is typical, as is the uncertainty as to the evidential weight to be accorded the various kinds of available information. As Helmer and Rescher (1959) suggest, "Hence the great importance which must be attached to experts and to expertise in these fields. For the expert has at his ready disposal a large store of (mostly inarticulated) background knowledge and a refined sensitivity to its relevance, through the intuitive application of which he is often able to produce trustworthy personal probabilities, regarding hypotheses in his area of expertness."

As an example of the use of experts in an environmental engineering context, a study undertaken to develop a plan for the management of solid wastes in a typical urban-agricultural region of California might be cited (Fresno Region Solid Waste Management Study, 1969). Experts were used to identify the environmental effects of solid waste and the conditions under which solid waste was likely to occur in the study area. In another phase of this project, experts in the engineering and environmental health fields provided value judgments as to the relative contribution to possible undesirable effects of a given waste under a given condition. They also quantitatively scored the relative importance of the various undesirable effects in terms of the type of area or subregion in which they occurred and determined the relative contribution to the generation of undesirable effects by solid waste as compared to other contributors. In short, the use of experts comprised a major portion of this study. This, however, is typical of complex problems in the environmental and engineering sciences, problems in which a great deal of uncertainty is encountered.

Some of the important questions addressed in this review are: What criteria can be established to evaluate experts? Do we obtain better results from experts if they themselves combine the various judgmental components that enter into an overall or global judgment, or if these components are combined in some mechanical fashion to form the global judgment? If the latter is the case, what models can be suggested to combine the components? If more than one expert is utilized, how may we go about achieving a consensus? Are some techniques for achieving consensus superior to others? When experts are required to produce judgmental probability distributions, as opposed to a single deterministic estimate, what modifications of our techniques must be made? This review critically explores the progress made to date in attempts to answer these questions.

CONCLUSIONS AND SIGNIFICANCE

Section I discusses two criteria for evaluating experts: reliability and accuracy. The realities of implementing such criteria, however, are such that one cannot claim absolute or objective validity for their use when past experience is meager or lacking.

The combination problem (Section II) deals with how the individual components of an expert's judgment are combined to form an overall or global judgment. The evidence suggests that mechanical combination (for example, the utilization of a model whose parameters are fitted by

regression analysis) is superior to combination by the expert. Several combining models are discussed: linear, conjunctive, and disjunctive. Although it is not possible a priori to state which is superior in a given case, enough is known about what each model attempts to do to enable the decision-maker to make a selection on the basis of the objectives of the problem under consideration. The decision-maker may, however, encounter psychological problems in using the inputs of experts in a mechanical or actuarial nature in that experts usually like to be a part of the overall decision or judgment and may resent being used in this fashion.

Two basic approaches to achieving a consensus when two or more experts are used are mathematical and behavioral. The former is discussed in Section III, the latter in Section IV. A number of weighting schemes are available for achieving a consensus using a mathematical approach. In general, the evidence indicates that a true consensus (that is, one based upon components rather than global judgments) is superior to a simple averaging of the global judgments of the individual experts. Since some experts may be more accurate with regard to certain components than others, superior results may be achieved by differentially weighting experts according to those components they estimate best.

Traditional behavioral approaches (for example, meeting as a group, conducting discussion, and ultimately arriving at a group consensus) have been plagued with undesirable psychological factors and game-theoretic strategies dysfunctional in arriving at a valid consensus. One remedy is the Delphi approach, a form of feedback and reassessment technique that avoids these undesirable features. The Delphi technique is cumbersome, however, and difficult to implement when quantitative statements are required from the experts. Behavioral approaches do lead to the consideration of factors which might otherwise be ignored and also reduce the spread or variance of the experts' judgments when quantitative estimates are required.

When the decision-maker requires judgmental probability distributions from two or more experts, a consensus can be obtained under certain circumstances by the application of Bayes' Theorem (Section V). The advantage to this approach is that the consensus distribution so obtained has smaller uncertainty and is generally unimodal in nature. The procedure is applicable, however, only when the requirements of Bayes' Theorem are congruent with the true judgments of the experts. Since the sort of probability distributions accommodated by the Bayes' Theorem approach are extremely flexible in shape and form, the method should have broad applicability. Several scoring rules for obliging the expert to reveal his true judgments are discussed. The selection of such a rule, however, is not only a matter of the mathematical theory involved, but of communication between the expert and the decision maker as well.

Section VI briefly considers the question of how many experts should be used in important problems requiring consensus. For mathematical approaches, little improvement is noted when more than 5 experts are utilized. In behavioral approaches, the optimal group size is between 8 and 12. More research needs to be done in this area, however.

The final technical Section (VII) deals with the expert's use of the decision maker, having particular relevance to game theoretic strategies that may be used by engineering consulting firms. It can be shown that there are strategies that can be used by the expert to maximize his goal achievement, expressed in the form of the probability that his publically stated advice will be accepted by the decision maker.

I. CRITERIA FOR EVALUATING EXPERTS

What constitutes an expert is an important question. The first criterion is the obvious one of knowledge, but this alone is not sufficient for he must be able to bring that knowledge to bear effectively on the problem. The simplest way in which to score an expert's performance is in terms of reliability, which may be defined as the relative frequency of cases in which, when confronted with several alternatives, he ascribed to the eventually best alternative a greater preference than to the others. In the area of decision-making in general, the decisions themselves inevitably turn on the question of future developments since present actions are almost invariably conceived with a view to future results. Hence, experts are often sought for their predictive ability. Reliability in this sense, then, is the relative frequency in which the expert ascribed to the actual event a greater probability than to the other events.

The reliability criterion, however, may be misleading for at times even laymen may attain high reliability. If about 80% of the days in a particular area experience good weather, and if one always predicts a fine day, then one will have an excellent reliability without being an expert in the meteorological sense. What is important is relative reliability, not absolute reliability, where relative reliability is defined (crudely) as reliability as compared to that of the average person. (Even this may not be sufficient, for the average layman may not be aware that 80% of the days are fair. Yet, an informed layman would still not be regarded as an expert. Thus, the definition of relative reliability might better include a comparison with the average person having some degree of general background in the required specialization.) A more subtle measure of expert performance is the degree of accuracy of his predictions. This is the correlation between the expert's probabilities and what is actually observed. For example, if an expert ascribes a probability of 70% to an event under certain circumstances, then under the given circumstances the event should occur 70% of the time. Accuracy, in this sense, does not guarantee reliability. Consider two experts, A and B, who are asked to make a prediction about each of 100 cases. Both ascribe a probability of 60% (that the prediction is right) to each of their 100 estimates. Suppose that 60% of A's estimates and 100% of B's estimates are correct. Then A is perfectly accurate since exactly 60% of the cases he predicted and to which he assigned a probability of 60% were correct.. He is, however, only 60% reliable. B, on the other hand, is perfectly reliable but only 60% accurate. Although accuracy does not guarantee reliability, the real mark of the expert may be the combination of accuracy with reliability. The realities of implementing rules such as this one, however, are brought to us nicely by Savage (1971):

> "No rule of this sort can claim absolute or objective validity. For example, actual past experience with the

experts may be extensive, moderate, meager, or absolutely lacking. When past experience is extensive... the rule often has much to recommend it; when direct past experience with the experts is meager, the rule is silly; and when such experience is altogether lacking, the rule results in a tie and is therefore empty. Actually, if you have little or no past experience with the experts, you will have to ponder them in terms of whatever information it was that brought you to regard them as promising in the first place: this well finder is regarded by the whole neighborhood as infallible with the hazel fork; that one is a professor of geology and the author of an important treatise on subsurface hydrology but has never before tried to help anyone locate a well. In such a context, the subjective aspect of your decision is thrown into prominence, but no matter how much direct past experience you may have with the experts, the ultimate subjectivity of your choice among them never disappears, though its effects may become less agonizing...."

Finally it ought not to go unsaid that experts have often been wrong in the past. Thomas A. Edison, writing in 1889, said: "There is no plea which will justify the use of high-tension and alternating currents, either in a scientific or a commercial sense." In a speech in the U.S. Senate, Senator Morris Sheppard, author of the 18th Amendment (Prohibition) commented: "There is as much chance of repealing the 18th Amendment as there is for a hummingbird to fly to the planet Mars with the Washington Monument tied to its tail." Hegel published his proof that there could be no more than seven planets just a week before the discovery of the eighth. In 1964, a panel of the National Academy of Sciences on Weather and Climate Modification called rainmaking by cloudseeding astrology. In 1966, the same panel had to admit that there was something to rainmaking after all. It took 20 years from the invention of cloudseeding in 1946 for a consensus to be reached among scientists that "it works!" In the interim, most of the advances that took place in the field occurred in spite of expert opinion.

II. THE COMBINATION PROBLEM

Most decision-making in the inexact sciences involves finding the utility of a multidimensional stimulus, that is, the decision-maker's task is to consider a number of inputs (components) and to weight or combine them in some fashion in order to arrive at an overall value or global judgment. The question arises, "Do we obtain better results from experts if the experts themselves combine the judgmental components they have developed to form a global judgment or the judgmental components are combined by the decision-maker in some mechanical or actuarial fashion?" The experimental evidence strongly suggests that the mechanical or actuarial mode of combination is superior to the expert or clinical mode of combination. A number of studies could be cited, but a recent one by Einhorn (1972) illustrates most of the major points to be made here.

The study dealt with the diagnosis of Hodgkin's disease (a form of cancer of the lymph system) by three highly trained pathologists. The basic problem involved predicting the survival time (in months) of 193 patients suffering from this disease. The sample information for each patient consisted of a biopsy slide taken when the patient was admitted to the hospital. All of the patients had died, but the pathologists had no knowledge of this or of the patients other than the biopsy slides. The pathologists, for each slide, were required (independently) to give their judgment as to the relative amount of nine histological characteristics that they saw in each of the slides (the experts themselves picked out the histological characteristics that they thought were important and formed the scales on which each of the signs was measured.) In addition, they also had to make an overall or global judgment as to the severity of the disease, again in terms of a rating scale the pathologists had agreed upon.

The correlations of the global judgments to actual survival time are shown in the first column of Table 1. The correlations are essentially random as none reach statistical significance using $\alpha = 0.1$. (The low correlations in Table 1 reflect the difficulty inherent in such diagnoses, the fact that the patients received different kinds of treatments, and that the patients may have died from other causes.)

The most often used models to describe a multidimensional utility function have been the linear, conjunctive, disjunctive, and lexicographic models (Einhorn, 1970). The linear model assumes that a global judgment is a linear combination of the individual inputs. The conjunctive model requires minimum scores on each of the input variables. It is a multiple cutoff procedure rather than the linear compensatory procedure exemplified by the linear model. With the conjunctive model, an input stimulus will produce a high global score only if all of the values for each input are high. A low score on only one input variable or component produces a low overall score. (If, for example, the input variables represented resistances to contacting various fatal diseases, a low score on even one of them would still mean that the patient dies.) In the disjunctive model, a high score on at least one input variable produces a high global judgment score. In other words, the global judgment is based on the most favorable input. (In selecting players for a football team, for example, we might want someone who can kick or run or pass with a great deal of skill. Using this model, the judgment or scoring of a player would be based upon his best ability regardless of his other attributes.) In the lexicographic model, the input variables are first ordered on importance and a decision made on the basis of the most important variable. One proceeds to the next ordered variable only if there is a tie on the preceding variable. Unfortunately, this model cannot be given explicit mathematical representation and will not be considered further here. A brief but more formal discussion of the linear, conjunctive, and disjunctive

Table 1. Correlations (R^2) of Components and/or Global Judgment with Survival Time

Judge	Global judgment Alone	Components alone Linear	Components alone Conjunctive	Components alone Disjunctive	Components and global judgment Linear	Components and global judgment Conjunctive	Components and global judgment Disjunctive
1	0.000	0.143	0.166	0.185	0.149	0.167	0.193
2	0.012	0.066	0.060	0.088	0.112	0.084	0.126
3	0.019	0.091	0.098	0.111	0.125	0.128	0.145

models is provided in the Appendix.

The linear, conjunctive, and disjunctive models were next used to examine the relationship between survival time and the nine individual judgmental components. The constants of these models were fitted by least squares, and the resulting coefficients of determination (R^2) computed. These are shown in the middle section of Table 1. Regardless of the model used, the component approach was clearly superior to the global judgment approach in that much larger correlations were obtained for all judges. Although there was not much difference among the models, the disjunctive model provided the best fit. Adding the global judgment as an extra input variable to the component models did not appreciably improve the correlations for judges 1 and 3, although there was significant improvement for judge 2. In any event, even when the global judgment was combined with the components, the disjunctive model remained superior for all judges.

A number of reasons have been advanced to explain why, in general, from studies of this sort global judgments are inferior to mechanical or actuarial combination:

1. The expert may be leaving out important components.
2. The expert may be weighting the components in nonoptimal ways.
3. The expert may be using a rule or model for combining the components to arrive at a global judgment that differs from the optimal combining rule to predict the criterion of interest.

With regard to the last, we have noted that, in the relationship between survival time and the components, the optimal combining model was the disjunctive. A similar procedure was used to relate the components to the global judgment, the results being shown in Table 2.

In this case, the conjunctive model provided the highest correlation with the global judgment. Clearly, the judges were using a model for arriving at their global judgments which was not optimal with regard to the point in using experts in the first place, that is, predicting patient survival time.

Other studies, particularly those of Sawyer (1966) and Meehl (1954), confirm the suggestion that a mechanical or actuarial mode of combination is superior to expert combination. The implication for the decision-maker who uses experts is clear. However, there may be certain psychological problems in using experts in this manner. Experts usually like to be a part of the overall decision or judgment and may resent being used solely as inputs to a mechanical model.

Table 2. Correlation (R^2) of Components with Global Judgment

Judge	Linear	Conjunctive	Disjunctive
1	0.115	0.200	0.181
2	0.000	0.005	0.005
3	0.109	0.198	0.076

Table 3. Correlation (R^2) with Survival Time, Nonconsensus and Consensus Approaches

Judge 1	0.193
Judge 2	0.126
Judge 3	0.145
Consensus, equal weights	0.204
Consensus, best performance	0.314

III. THE CONSENSUS PROBLEM: MATHEMATICAL APPROACHES

A problem may be important enough to justify consulting more than one expert. A new problem then arises in such cases—how is the best joint use of these various expert judgments to be made? One set of approaches is to combine the individual global judgments by means of some sort of weighted average, for example,

1. *Equal weights*: In this case the decision-maker has no reason to think that there is much difference in ability among the experts, so he is willing to assign equal weights.
2. *Weights based upon previous performance*: This method may be subjective or objective, depending upon the quantity and quality of the hard data obtained from prior experience with the use of the experts.
3. *Weights proportional to a self-rating*: Have each expert rate himself on a scale from 1 to M where M is the highest rating and 1 the lowest. Then assign a weight to each expert proportional to his self-rating, where the constant of proportionality is such that the weights sum to one (this constant is obviously the reciprocal of the sum of the self-ratings). The rationale behind this rule is that although a person may be an expert in a given field, his expertise may vary from topic to topic within the field.

Techniques that focus solely on global judgments do not, however, develop a consensus in the sense intended here. The term is reserved for those processes that involve the combination or, in some cases, modification of the components as seen by the experts individually.

Suppose there are two components to an estimate, and that experts A and B have rated these (6, 2) and (4, 4), respectively. Further assume that the components are to be combined according to the disjunctive model $Y = (10 - X_1)(7 - X_2)$. The experts' global judgments then are $20 = (10 - 6)(7 - 2)$ and $18 = (10 - 4)(7 - 4)$, respectively. If we weight these global judgments equally, the weighted average global judgment is 19. On the other hand, if we weight the experts' component judgments equally, the weighted ratings are (5, 3), resulting in a consensus of $20 = (10 - 5)(7 - 3)$.

The question arises, then, which is superior—weighting the experts on a component or global judgment basis? The evidence seems to indicate the former. Again, although many studies could be cited, Einhorn's 1972 study is illustrative. Table 3 compares the correlations, with survival time, of the optimal combining models for each of the three judges separately (these are taken from the "Components + Global Judgment, disjunctive" column of Table 1) with two consensus models. The "Consensus equal weights" model is based upon the average judgment of the three pathologists, component by component, except where they widely disagreed (in such cases, the doctors met in a group to discuss their differences and came to some agreement. Thus it is in the main, but not strictly, an equal weights model.) This consensus is superior to even the best of the doctors, and therefore clearly superior to any sort of average of the global judgments.

The last model in Table 3, "Consensus, best performance," considers that each judge may only be an expert in a subset of the judgmental components. For example, some engineers may be better at assessing cost information, while others may be better at assessing process yields. Therefore, one can view the terms in a combining model two-dimensionally, that is, input judgments and judges,

and thus arrive at many new variables. If there are m components (variables) and n judges, one can redefine mn new input variables and fit the model accordingly. The procedure used for the "Consensus, best performance" model was as follows. The $10 \times 3 = 30$ ratings (9 components + global, 3 experts) served as inputs for linear, conjunctive, and disjunctive models. A stepwise regression program was used to estimate the parameters of the models, using survival time as the criterion variable. The program was halted at the 10th step so that all of the models in Table 3 would have the same number of terms. The optimal model was the disjunctive, and it is quite superior to the others in Table 3. Clearly, there is differential validity among the experts with respect to the individual components. Thus, in some cases, the decision-maker may wish to use different weightings for his experts according to the component under consideration.

IV. THE CONSENSUS PROBLEM: BEHAVIORAL APPROACHES

Two techniques for dealing with the consensus problem may be described as behavioral in nature (Winkler, 1968). The first, termed "Group Reassessment," involves both psychological factors and game-theoretic strategies. In brief, the experts meet as a group and discuss the matter, ultimately arriving at a group consensus. In general, individuals tend to shift their views after group discussion in the direction of greater risk-taking. Psychological aspects involved include specious persuasion by the member with the greatest supposed authority or even merely the loudest voice, the unwillingness to abandon publicly expressed opinions, and the bandwagon effect of majority opinion.

Contrary to popular opinion, group decisions are riskier than individual decisions, that is, individuals tend to shift their decision preferences after group discussion in the direction of greater risk-taking. Studies have shown that the risk-taking individual who is disproportionately influential in a group situation is extraverted, has a high need for achievement, and is tolerant of ambiguity; his theoretical, economic, and political interests are high, as are his interpersonal relations (Rim, 1966). Game-theoretic aspects must be considered also, for example, an expert may intentionally falsify his stated views in an attempt to influence the others. This is discussed in additional detail, from another point of view, in Section VII of this review. The major advantage to Group Reassessment is that it may lead to the consideration of factors which might otherwise be ignored.

As has been mentioned, the outcome of the Group Reassessment method is apt to be a compromise among divergent views, arrived at all too often under the undue influence of certain psychological factors. One way to reduce these effects is to allow the experts to discuss the situation without arriving at a group decision or consensus. This leads us, however, to the second of the behavioral methods of arriving at a consensus, namely, Feedback and Reassessment techniques. Here each expert, separately, reconsiders his assessments after being presented with some feedback regarding all of the experts. This process of feedback and reassessment can be carried out repeatedly.

An interesting version of the Feedback and Reassessment method is the Delphi technique (Helmer, 1967). In its simplest form, it eliminates committee activity altogether and replaces it with a carefully designed program of sequential individual interrogations (usually best conducted by questionnaires) interspersed with information and opinion feedback. The principles involved in this procedure are illustrated by the following example.

Each member of an engineering review panel was asked to examine a number of newly-submitted proposals and to rate them according to a pre-established scale. The initial responses consisted of these ratings. A follow-up questionnaire fed back to the panel members included a summary of the distribution of these ratings in the form of the median for each proposal and—as an indication of the spread of opinions—the interquartile range (that is, the interval containing the middle 50% of the responses). Each panel member was then asked to reconsider his previous ratings and revise them if he desired. If his new response lay outside the interquartile range for a particular proposal, he was asked to state his reason for thinking that the rating should be that much lower, or higher, than the majority judgment of the group. (In actual experience, placing the onus of justifying relatively extreme ratings on respondents has the effect of causing those without strong convictions to move their estimates closer to the median, while those who feel they have a good argument for a "deviationist" opinion tend to retain their original estimate and defend it.)

In the next round, responses (now spread over a smaller interval) were again summarized, and the panel members were given a concise summary of reasons presented in support of extreme positions. They were then asked to revise their second-round responses, taking the proffered reasons into consideration and giving them whatever weight they thought was justified. A panel member whose answer still remained outside the interquartile range was required to state why he was unpersuaded by the opposing argument. In a fourth and final round these criticisms of the reasons previously offered were resubmitted to the panel members, and they were given a last chance to revise their ratings. The median of these final ratings was then to be taken as representing the nearest thing to a group consensus. (Note: The Delphi procedure causes the interquartile range to shrink considerably, presumably influenced by convincing arguments. In those rarer cases where no convergence toward a relatively narrow interval of values takes place, opinions tend to polarize around several distinct values, so that several schools of thought regarding a particular issue seem to emerge. In any event, the Delphi technique serves the purpose of crystallizing the reasoning process that might help lead to one or several positions on an issue, and thus helps to clarify the issue even in the absence of a group consensus.)

It will be noted that the Delphi technique as used in this example was fairly simple to implement and to administer. Four questionnaires and three summaries were used as follows:

1. Questionnaire I — Initial Response
2. Summary A — Means and Interquartile Range (M & IR)
3. Questionnaire II — Reconsideration and Statement of Reasons
4. Summary B — M & IR plus Summary of Reasons
5. Questionnaire III — Reconsideration and Reactions to Opposing Reasons
6. Summary C — M & IR plus Summary of Reactions to Opposing Reasons
7. Questionnaire IV — Final Response

Since the questionnaires and summaries in this example

TABLE 4. DELPHI METHOD—FEEDBACK AND ITERATION

Question no.	Results of original questionnaire 25th percentile	50th (median)	75th percentile	Results of first iteration 25th percentile	50th (median)	75th percentile	Correct answer
1	150	5,000	1,000,000	500	3,000†	10,000	21,578
2	3	50	240	20	77*	240	110
3	100	982	10,000	500	5,000*	8,000	31,511
4	10	200	500	100	200	300	165
5	80	500	1,000	100	500	640	1,490
6	50	200	700	52	200	500	286
7	200	1,000	6,000	500	1,000	5,000	5,057
8	1,000	5,000	25,000	1,000	8,000	10,000	19,800
9	400,000	800,000	1,000,000	800,000	1,000,000	1,000,000	1,854,700
10	1,000	5,000	10,000	2,500	5,000	6,000	6,253

* More accurate.
Same.
† Less accurate.

were brief and concise, and the former easily filled out, considerable panel participant time was saved.

The illustration given above describes the basic essentials of the Delphi technique. There are refinements, of course, such as the introduction of weighted opinions, but the potentialities are clear. A recent demonstration of the Delphi method described the results of a series of graduate-engineering classroom exercises in formulating group judgments for which the correct answers were known (Doyon et al., 1971). The group was charged with answering ten questions as follows:

1. What was the number of telephones (in thousands) in Asia in 1967?
2. How many million barrels of beer were produced in the U.S. in 1966?
3. How many millions of dollars were expended for public elementary and secondary education in the U.S. in 1967?
4. What was the total U.S. Budget income in 1968 (in billions of dollars)?
5. How many thousand new private housing units were started in the U.S. in 1969?
6. In 1967 what was the average number of pounds of milk consumption per person in the U.S.?
7. What was the total motor fuel consumption in millions of gallons in 1967 in N. Y. State?
8. What was the total number of deaths from falls in the U.S. in 1967?
9. How many elementary and secondary school teachers were there in the U.S. in 1967?
10. How many radio and TV broadcast stations were operating in the U.S. in 1967?

Table 4 summarizes the results after only one round (that is, one questionnaire and feedback), indicating that not only did the accuracy of the estimates improve, but that the variance within the estimates decreased.

The principal drawback of the Delphi technique would appear that it is cumbersome (English, 1968). Obviously, time must elapse while questionnaires are prepared, responses gathered, reconsiderations made, and summaries prepared. Some thought has been given, however, to computerizing the process. The U.S. Navy has utilized an electronic device that permits the anonymity of the experts but not their reasons. At the Rand Corporation, a system of personal electric typewriters connected through an on-line computer has been used to eliminate both questionnaires and middlemen. However, other deficiencies remain. Leading by those conducting the Delphi processing may influence the results. Also, it is difficult to summarize or condense qualitative statements, or indeed even to define an interquartile range or variance for such material.

Winkler (1972) describes an experiment in which respondents were asked questions about which they were expected to have some knowledge (the questions covered such topics as sporting events, weather, and stock prices) but about which there was some uncertainty, obtaining consensus using equal weights, self-rating weighted average, feedback and reassessment, and group reassessment methods. The differences among these experts were reduced after applying group reassessment and feedback/reassessment, with smaller differences noted for the former. The consensus obtained by either behavioral reassessment method differed much more from either of the mathematical averaging methods than did the two averaging methods from one another, supporting the general conclusion of this and other studies that consensus is not affected as much by the choice of weighting schemes as by the choice of methods.

V. JUDGMENTAL PROBABILITY

Estimates involve uncertainty, and rather than settling for a single input number, the decision-maker may require the expert to produce a judgmental probability distribution that reflects this uncertainty. If so, then yet another way to combine the experts' input is by means of natural-conjugate distributions. The concept is rooted in Bayes' Theorem which, under certain conditions, suggests that an expert's judgment can be thought of as being equivalent to sample information from the data-generating process of interest. The natural-conjugate approach has been developed in an attempt to simplify Bayesian analysis, and although the mathematical development of Bayes' Theorem and natural conjugates cannot be described here (see Hadley, 1967), a few results can be presented.

For example, if the experts' judgmental probability distributions can be considered to be normal with means m_i and variances v_i (the subscript i signifying a particular expert), and if these individual normal distributions are weighted and combined via Bayes' Theorem, the consensus would be a normal distribution with mean $m = \Sigma(w_i m_i/v_i)/\Sigma(w_i/v_i)$ and variance $v = 1/\Sigma(w_i/v_i)$, where w_i is the weight applied to the ith expert. As another example, if the experts' judgmental probability distributions

were beta with parameters r_i and n_i, the consensus would also be beta with parameters $r = \Sigma w_i r_i$ and $n = \Sigma w_i n_i$. In the latter case, the consensus consists simply of expert-by-expert addition of the two parameters.

In utilizing the natural-conjugate method, although there is no theoretical restriction on the sum of the weights, if the judgments of k experts are completely independent in the sense that their judgments are based upon independent sets of information (such as independent samples), then it seems reasonable to set each weight equal to one so that the sum of the weights equals k. On the other hand, if the judgments of the experts are based on exactly the same information (such as the same sample) the sum reasonably should equal one. This suggests placing a lower bound on the sum of the weights of one, and an upper bound on k.

Suppose, for example, that expert A considers that his uncertainty of the market price for a future commodity is adequately described by a normal distribution with mean $m_1 = \$2$ and standard deviation $\sqrt{v_1} = \$1$, and that expert B feels his uncertainty is best reflected by a normal distribution with mean $m_2 = \$3$ and standard deviation $\sqrt{v_2} = \$2$. We will assume that the experts' opinions are based upon strictly independent information so that $w_1 + w_2 = 2$, and that we consider each opinion to be equally important, that is, $w_1 = w_2$. Thus, $w_1 = w_2 = 1$. Using the results described above, the consensus distribution would be normal with mean $m = (2/1 + 3/4)/(1 + 1/4) = \$2.2$ and standard deviation $\sqrt{v} = \sqrt{1/(1/1 + 1/4)} = \$2.8$.

One could, of course, dispense with Bayes' Theorem and the natural conjugate approach, and simply compute a weighted average of the individual distributions (where $\Sigma w_i = 1$). There are two main differences, however, between a weighted average of judgmental probability distributions and the natural conjugate method. First, the weighted average method produces a combined distribution which is more spread out. In this method, the tighter distribution does not, regardless of the choice of weights, automatically receive the most weight as it does in the natural conjugate approach. Second, the weighted average approach tends to produce multimodal combinations, even if the individual distributions are unimodal. The larger the difference between means and the tighter the individual distributions, the more likely is it that a bimodal distribution will result. The natural conjugate approach, on the other hand, in general produces unimodal distributions. It is, of course, a valid approach only if the probability distributions that truly express the judgments of the experts are, in fact, natural conjugates. However, many natural conjugate distributions are extremely flexible in form and shape, depending upon the selection of the parameters of the distribution (for example, the gamma and beta distributions), or occur so often in actual situations (for example, the normal distribution), that one would expect the technique to be applicable in a good many cases.

Noted earlier in this review were the problems encountered in establishing criteria for the evaluation of experts in the deterministic or nonstochastic case. With regard to scoring rules for probability assessments, there are many which seem reasonable but which turn out not to be upon further investigation. Consider a discrete probability distribution where $p_1, p_2 \ldots, p_n$ are the true probabilities for the n values of the random variable, and $r_1, r_2, \ldots, r_n$ are the expert's assessed probabilities. Let the subscript h refer to the interval in which the random variable is actually observed. An intuitively reasonable scoring rule is to let $S = r_h$. It turns out, however, that the expert can maximize his expected score by setting one of his r_i equal to one, that is, the r_i equal to the highest p_i, and the other r_i equal to zero. Suppose in a dichotomous situation, $p_1 = 0.6$ and $p_2 = 0.4$. By setting $r_1 = 1$ and $r_2 = 0$, the expert's expected score is $1.0\ (0.6) + 0.0\ (0.4) = 0.60$, whereas if he had stated what he actually believed (and assuming the expert is perfectly accurate), he would have stated $r_1 = 0.6$ and $r_2 = 0.4$, producing a score of $0.6\ (.6) + 0.4\ (.4) = 0.52$.

What is needed, then, is a scoring rule which will oblige the expert to reveal his true judgments. It can be shown (Winkler, 1967) that three rules that do this are:

(a) Quadratic scoring rule

$$Q = 2 r_h - \sum_{j=1}^{n} r_j^2,$$

(b) Spherical gain scoring rule

$$S = r_h \Big/ \left[\sum_{j=1}^{n} r_j^2 \right]^{1/2},$$

(c) Logarithmic loss scoring rule

$$L = \ln r_h.$$

The logarithmic rule clearly is the only scoring rule depending solely on the probability assigned to the event which actually occurs. It can be shown, therefore, to be the only scoring rule which may be used to keep the expert honest and to evaluate the expert (Winkler, 1969). The selection of an optimal scoring rule is not all cut-and-dried, however. Some scoring rules work better for some experts than for others in the sense that the rule helps them to become better assessors. People do not generally, for example, behave as Baysians, even if it is in their interest to do so. Further, the expert must understand the particular scoring rule employed, and the rule itself should encourage the expert to work hardest at what the decision-maker wants to know. The selection of a scoring rule, therefore, is a matter of communication, subject to experimentation.

An interesting study that compares several weighting schemes with the Bayesian method for the combination of probability distributions, involves the personal assessments of the outcome of football games (Winkler, 1971). The results are shown in Table 5.

The first three lines of data are not consensus scores, but are included for comparison purposes. The first two scoring rules, Quadratic and Spherical, show all of the con-

Table 5. Results of Hypothetical Bets for 1967 NFL Games

Weighting Method	Scoring Rules: Quadratic	Spherical	Logarithmic
Average of all subjects*	57.3	0.41	−2.07
Worst of all subjects	53.9	0.35	−2.80
Best of all subject	59.9	0.45	−1.18
Equal weights	59.6	0.44	−1.70
Weights from self-ratings	59.6	0.44	−1.70
Weights based upon past performance**	59.4	0.44	−1.67
Bayesian method	53.1	0.39	−1.50

* The number of subjects was 10.
** Data from 1966 games.

TABLE 6. AVG. QUADRATIC CONSENSUS SCORES FOR 1967 NFL GAMES AS A FUNCTION OF SIZE OF GROUP COMPRISING THE CONSENSUS

No. of individuals	Equal weights	Self-rating weights
1	57.3	57.3
2	58.6	58.5
3	59.0	58.9
4	59.2	59.2
5	59.3	59.3
6	59.6	59.6
7	59.5	59.5
8	59.5	59.5
9	59.5	59.6
10	59.6	59.6

sensus methods, with the exception of the Baysian method, to be superior to the average of all subjects, and to match closely the performance of the best of all subjects. There was little difference among the first three consensus systems. Thus, a simple weighted average outperformed most of the individuals comprising the consensus. For the logarithmic scoring rule, the Bayesian method performed best, but this is not surprising since it is the only rule consistent with the use of the Bayesian model used to evaluate experts. In any event, the point to be made here is that, unless the decision-maker feels very confident that he can identify in advance the few experts who will "beat the consensus" in any given situation, he will be better off using a consensus.

VI. HOW MANY EXPERTS TO USE?

For the football scores study just described, all possible combinations of k assessors were considered and averaged for each k, $k = 1, \ldots, 10$. (Therefore, $k = 1$ corresponds to a simple average of the scores of the 10 individual assessors, and $k = 10$ corresponds to the consensus of all 10 subjects.) The results are shown in Table 6 for both equal weights and self-rating weights. (Only the average Quadratic scores are shown, but the results for the Spherical and Logarithmic scoring rules are similar.)

The scores rise rapidly as k is increased, and level off near $k = 5$. In Delphi studies, on the other hand, it has been found that the average group error drops rapidly as the number in the Delphi group reaches eight to twelve, at which time very little improvement it attained by increasing the group size (Fusfeld et al., 1971).

VII. THE STRATEGY OF ADVICE

Although this review primarily has been concerned with the decision-maker's use of experts, a word might be said about the expert's use of decision-makers. This is especially relevant when the expert is in the role of an advisor where we may assume that he desires to attain goals of his own choosing. The expert, however, will not necessarily come closest to his own goal by stating it and arguing for it. If his advice is at all extreme from the decision-maker's point of view, he may find his recommendation written off as unrealistic. The expert's effectiveness and degree of goal achievement will then be low or even zero. On the other hand, if in order to sound realistic and maximize

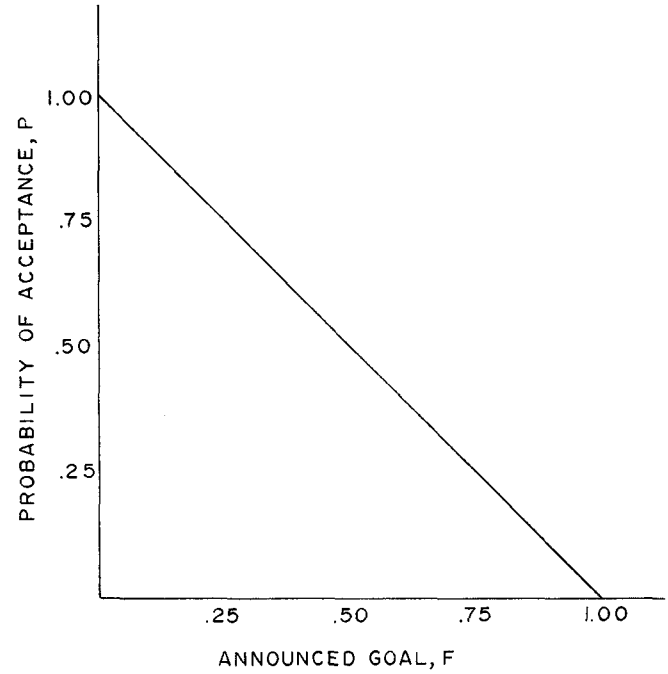

Fig. 1. Probability P that the expert's advice will be accepted versus announced goal F expressed as a fraction of the difference between the decision-maker's views and the expert's true goal.

the probability of having his advice accepted the expert shades his goal to the point of making it the same as that of the decision-maker he serves, his degree of goal achievement, expressed as the difference between the decision-maker's views with and without the benefit of his advice respectively, is also zero. In between, there is a maximum expected goal achievement. To illustrate this point, let us assume that the functional relationship between the probability of the expert's advice being accepted and the proportion by which he understates his true goal is linear. Further assume that the only possibilities are (a) acceptance of the expert's announced goal and (b) total rejection and that statement of the expert's goal will reduce its probability of acceptance to zero. This situation is shown graphically in Figure 1. The equation of this function is

$$P = 1 - F.$$

The expected (average) goal achievement of the expert is his announced goal (that is, advice) F, times the probability that this advice will be accepted. In other words,

$$FP = F(1 - F).$$

By elementary calculus, we maximize FP by taking the first derivative of this function and setting it equal to zero, that is,

$$d(FP)/dF = 1 - 2F = 0.$$

Thus $F = 1/2$, that is, the expert will maximize his goal achievement if he makes his announced goal equal to one-half of the difference between his true goal and that of the decision-maker he serves. Different maxima will result if the expert, because of his greater or lesser powers of persuasion, can assume the relation between goal announcement and probability of acceptance to be, for example, concave or convex (Wallich, 1968). The implications for engineering consulting firms and for those who utilize such firms is that consultant might not always reveal his true judgments, even if he is paid for doing so.

VIII. SUMMARY

Our present knowledge of the best techniques and procedures for the selection, evaluation, and utilization of experts is, admittedly, far from perfect, but important advances are being made. The role that experts play in complex decision processes is a large and growing one, however, and we can expect that attempts will continue to be made to use them even more effectively in the future.

LITERATURE CITED

Doyon, L. R., T. V. Sheehan, and H. I. Zagor, "Classroom Exercises Is Applying the Delphi Method for Decision-Making," *Socio-Econ. Plan. Sci.*, **5**, 363 (1971).

Edison, T. A., "The Dangers of Electric Lighting," *North Amer. Rev.*, **75**, 82 (1889).

Einhorn, H. J., "Expert Measurement and Mechanical Combination," *Organiz. Behavior Human Performance*, **7**, 86 (1972).

———, "The Use of Nonlinear, Noncompensatory Models In Decision Making," *Psychol. Bull.*, **73**, 221 (1970).

Engineering-Science Inc./Aerojet-General Corp., "Fresno Region Solid Waste Management Study," Standard Agreement No. 15100, State of Calif., Berkeley (1969).

English, J., *Cost Effectiveness*, pp. 252-253, Wiley, New York (1968).

Fusfeld, A. R., and R. N. Foster, "The Delphi Technique: Survey and Comment," *Business Horizons*, 14, No. 3, 63 (1971).

Hadley, G., *Introduction to Probability and Statistical Decision Theory*, Holden-Day, San Francisco (1967).

Helmer, O., "Analysis of the Future: The Delphi Method," pp. 1-11, Defense Doc. Center Document AD 649640, The Rand Corp. (1967).

———, and N. Rescher, "On the Epistemology of the Inexact Sciences," *Management Sci.*, **6**, 25 (1959).

Meehl, P. E., "Clinical vs. Statistical Prediction," Univ. Minnesota Press, Minneapolis (1954).

Rim, Y., "Who Are the Risk-Takers in Decision-Making?" *Personnel Admin.*, **29**, 26 (1966).

Savage, L. J., "Elicitation of Personal Probabilities and Expectations," *J. Am. Statistical Assn.*, **66**, 783 (1971).

Sawyer, J., "Measurement and Prediction: Clinical and Statistical," *Psychol. Bull.*, **66**, 178 (1966).

Wallich, H. C., "The American Council of Economic Advisors and the German Sachverstaendigenrat: A Study on the Economics of Advice," *The Quarterly J. Economics*, **82**, 350 (1968).

Winkler, R. L., "The Consensus of Subjective Probability Distributions," *Management Sci.*, **15**, 61 (1968).

———, and L. L. Cummings, "On the Choice of a Consensus Distribution in Bayesian Analysis," *Organiz. Behavior Human Performance*, **7**, 63 (1972).

———, "The Quantification of Judgment: Some Methodological Suggestions," *J. Am. Statistical Assn.*, **62**, 1105 (1967).

———, "Scoring Rules and the Evaluation of Probability Assessors," *ibid.*, 1073 (1969).

———, "Probabilistic Prediction: Some Experimental Results," *ibid.*, **66**, 675 (1971).

APPENDIX: A SUMMARY OF SOME COMBINATION MODELS

A linear model may be represented by

$$U = \sum_{i=1}^{n} w_i X_i \qquad (1)$$

where U is the overall or global judgment represented in some quantitative form, and w_i are the weights given to the individual inputs X_i. (The form of this model, as well as those that follow, as it involves addition and multiplication, implies that a cardinal scale is being employed.) Such a model, of course, does not provide for interactions or nonlinear relationships. Linear regression models containing terms such as X_i^2 and X_iX_j, however, obviate this difficulty and have proved extremely effective in predicting actual global judgments from component cues.

There are a number of functions that can be used to approximate the conjunctive model; the parabolic function is one of the simplest:

$$U = \prod_{i=1}^{n} X_i^{w_i} . \qquad (2)$$

Clearly, a low value for any X_i will produce a low value for U.

As in the case of the conjunctive model, a number of functions approximately meet the disjunctive requirement but a simple one is the hyperbolic function,

$$U = \prod_{i=1}^{n} (a_i - X_i)^{-w_i} \qquad (3)$$

where the a_i are some values greater than the maximum possible X_i scores (to prevent the value of U from becoming infinite). A high value for any X_i will produce a high value of U in this model.

If the weights in the conjunctive and disjunctive models are to be determined experimentally, a preliminary step can be to take logarithms of both sides of the equations, namely,

$$\ln U = \sum_{i=1}^{n} w_i \ln X_i \qquad (4)$$

and

$$\ln U = -\sum_{i=1}^{n} w_i \ln(a_i - X_i) \qquad (5)$$

and fit Equations (4) and (5) by ordinary linear regression methods.

THE AUTHOR

Albert J. Klee is Deputy Director of the U.S. Environmental Protection Agency's Solid Waste Research Laboratory, and also Adjunct Associate Professor of Mathematics and Management at Xavier University, Cincinnati, Ohio. He holds the following degrees: B. Chem. Eng. (City College, New York, 1950), M. Chem. Eng. (New York University, 1955), M.B.A., (Xavier, 1959), and M.S. in Mathematics (Xavier, 1962). He is currently completing requirements for an interdisciplinary doctorate in systems analysis, economics, and public administration at the University of Cincinnati. Klee is the author of approximately 30 technical papers which have appeared in such journals as *Journal of the Sanitary Engineering Division of the American Society of Civil Engineers, American City, Environmental Science and Technology, American Public Works Reporter, Geoscience Electronics, Journal of the American Institute of Chemical Engineers,* and *Management Science.* His primary research interests have focused on computer modeling, decision theory, and attitude arousal, change, and measurement—with reference to environmental problems. Klee is a part-time explorer and amateur ichthyologist of some standing. He has explored the jungle areas of Peru, Ecuador, Guatemala, Honduras, Brazil, and Columbia, among other countries, and has discovered two new genera of fishes and several new species. A species of fish—*apistogramma Kleei*—found in the Amazon Basin is named after him.