T-SMC/18/1//18011

Sources of and Effects of Interexpert Correlation: An Empirical Study

Jane M. Booker
Mary A. Meyer

Reprinted from
IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS
Vol. 18, No. 1, January/February 1988

Correspondence

Sources of and Effects of Interexpert Correlation: An Empirical Study

JANE M. BOOKER AND MARY A. MEYER

Abstract —**Expert estimates are relied upon as sources of data whenever experimental data are lacking, such as in risk analyses. Correlation between experts poses problems in the aggregation of multiple estimates. The sources and structure of interexpert correlation are identified, and some of the ramifications are discussed.**

I. Introduction

This study seeks to identify sources of correlation in expert estimates to provide useful information to decisionmakers and others who work with expert judgments. Correlation is defined here to mean dependence. Correlated estimates are considered to be those where the experts have arrived at similar estimates because of some condition held in common, such as training.

The identification of sources of correlation is important to solving problems associated with expert opinion, such as how to aggregate multiple estimates. Most aggregation schemes require that the data be independent, even though experts' estimates are not known to be. Thus the analyst is led to assume independence to aggregate the data and to proceed with the analysis. This correspondence offers a means for aggregating multiple correlated estimates.

It focuses on an area for which little data and a wealth of speculation existed. There had been no studies on what might cause in<span>ı</span>erexpert correlation. Furthermore, little data from real settings existed that could be used to investigate possible sources because data on the experts had not been gathered. Despite the dearth of data on sources, the tacit assumption of many was that expert data were correlated. Several conditions were speculated to be the cause of agreement in the experts' estimates. Due to this situation of sparse data and abundant speculation, an empirical field study was conducted rather than a simulation study.

Section II elaborates on some of the possible sources of interexpert correlation. The study is described in Section III, including the questionnaire design and method of interviewing. We interviewed 18 statisticians, asking them statistical questions of the types encountered in walk-in consulting situations. The techniques for analysis of the collected interview data and the results are given in Section IV. Conclusions and recommendations are presented in the final section.

II. Background

A. The Search for Possible Sources of Correlation

Although expert opinion is commonly used as data, the question of correlation between experts [1], much less sources of correlation, has not been studied. Yet there has been much speculation and the making of tacit assumptions. For example, we have heard analysts in the expert opinion field state that they believed expert estimates to be correlated.

Sources of correlation have also been assumed without the basis of any real evidence. One speculation is that the individuals' shared work experiences or exposures to the same databases may lead to correlation. Another speculation is that the theoretical orientations individuals learn during their education may be a source of correlated estimates. For example, Baecher [2] was able to group seismic experts into three groups according to their answers on hazard forecasting. He labeled the groupings "schools of thought" but could not determine what "schools" these were because the experts had been anonymous.

Because literature on causes of interexpert correlation was lacking, we did some of our own speculation and reviewed literature in such areas as the factors influencing individuals' problem solving. Some studies propose that traits which humans have in common may influence their problem solving. In particular, humans possess memory limitations [3] that affect their reconstruction of events and their equations for solving problems. For this reason the recency of the subject's experience in working a similar problem was postulated as a source of correlation.

A number of studies had shown that the problem itself has an effect on the answers given. For example, Tversky and Kahneman [4] have shown that the presentation of the decision task influences the individual's response. The relative attractiveness of options varies when the same decision problem is framed in different ways. We designed the technical questions asked of our statistical experts using different formulations to counteract bias. For example, some questions were phrased in the usual problem format, and some were phrased in an unfamiliar reversed fashion. It was not the intent of this study to test the effects of different question formulations as a correlation source.

In another study, the complexity of the questions seemed to have some effect on the answers [5]. There seemed to be an increase in variance between answers for those questions that had been rated as more complex. These studies demonstrate a link between the question itself and the answers that are selected. For this reason the statistics questions were designed to have differing complexities that could be examined for their effect on the subjects' responses.

Some studies have shown that the individual's problem-solving techniques influence his answer. For example, if individuals are instructed or assisted in breaking a problem into its component parts and in solving the parts, they give more accurate ans.<span>/</span>ers than do those who have not used this problem-solving technique [6]. In general, the heuristics that individuals use to reduce the mental effort of solving the problems bias their answers in some manner [3]. Matz [7] examines individuals' problem-solving errors in her attempt to understand their solutions. She views the errors as stemming from reasonable but unsuccessful attempts to adopt old knowledge to new situations. The assumptions that individuals make are also likely to influence their answers. Ascher [8] determined that one of the major sources of inaccuracy in forecasting future possibilities, such as markets for utilities, lay in the forecaster's failure to extrapolate sufficiently from present patterns. The experts assumed that the present situation was likely to exist in the future, and when the future became the present, the forecast was proved incorrect because of this assumption. Based on these studies, the decision was made to carefully monitor the assumptions made by the subjects.

The studies just mentioned hinted that the ways in which individuals solved problems might be a source of correlation. The decision was made to record in detail the steps individuals took in answering the questions.

Manuscript received March 3, 1986; revised September 19, 1987. This work was supported in part by the Division of Risk Analysis, Office of Nuclear Regulatory Research, Nuclear Regulatory Commission, under FIN A7225. This correspondence was presented in part at The Joint Statistical Meetings, Las Vegas, NV, August 1985.

The authors are with the Statistics Group (F600) of Los Alamos National Laboratory, Los Alamos, NM 87545.

IEEE Log Number 8718011.

In summary, this study tested the following sources of correlation proposed in the studies just cited: common educational background and common work experience. In addition, we proposed and tested two other sources: the process the individual used to solve problems, and the length of time since the individual had worked on a similar problem. Also, two factors were tested for their effect on the individual reaching the "correct" answer: the complexity of the problem and the individual's problem-solving technique.

B. Aggregation and the Correlation Controversy

A few authors have acted on the belief that significant correlations exist and have investigated means for aggregating expert estimates. Aggregation is important for two reasons: it is convenient to have a single value representing the total knowledge of the experts, and it is assumed that such a pooled value will provide the best estimate of the true (unknown) value. In a recent study by Martz *et al.* [9], several pooling methods for forming aggregation estimates were compared in an experimental study. His paper also provides information on other references relating to the various pooling techniques. However, in any aggregation technique the assumption of independence of the data is usually required. Correlated estimates violate this assumption. Correlation then becomes an analyst's nightmare. This is part of the correlation controversy—correlated data are undesireable because they are difficult to analyze. For this reason many analysts make the assumption that the experts' estimates are *not* correlated.

Winkler [10] provides methods for handling the dependencies in the data from correlation by examining estimation errors using a Bayesian context. Lindley and Singpurwalla [11] also discuss correlation effects in Bayesian formulation of reliability and fault-tree applications. These analytic solutions assume that correlation exists and that its structure is somehow known or estimable.

In spite of the analytic problems, some authors seem to interpret correlation as a positive trait and others as a negative one. Some view correlation as desirable, as an indication of consistency, reliability, and accuracy. Disagreement between the experts is viewed as the contrary. They tend to eliminate outliers and value homogeneous expert populations [5]. They may also find positive merit in not being able to link the experts' estimates to their age, experience, or organization [12].

Expert correlation is tacitly viewed as negative by those who place a high value on the diversity of opinions. For example, correlation could be viewed as evidence that the problem has not been considered from enough perspectives to obtain a quality answer [13], that the experts are making one major and probably highly conservative assumption [8], or that the experts are unconsciously following one person's view [14]. Outliers are not eliminated under this interpretation because diversity, rather than consistency, is valued. This work does not place a positive or negative value on interexpert correlation but simply seeks to understand its causes.

III. Interviewing Method and Questionnaire Design

The lack of information on sources of correlation, the fuzziness of human-generated data, and the difficulty in discovering the subtle causes of correlation drove the design of the experiment in three ways: 1) test situation, 2) intensive interviewing, and 3) data-driven analysis.

A. Test Situation

In applications where expert estimates are sought, the problem areas range from those where the estimates could be verified, to problem areas (such as some forecasting) where estimates cannot feasibly be verified We chose a population of statisticians whose expertise was familiar to us so that the information they provided could be easily understood. This was important since much of the information being elicited was on problem-solving processes. A strategy of asking short multiple technical questions was chosen given the one hour time limit for the interview. A multiple question design has the advantage over a single lengthy question design in that it provides more opportunities to discover and to test for the subtle effects of correlation being sought.

Because the interviews were limited to one hour, subjects were not allowed to use any tools of the trade (tables, calculators, or reference texts). Since the technical questions covered commonly encountered walk-in problems, we did not feel that this hampered the subjects ability. Instead, it may have served to tap the expert's intuition and experience in problem solving.

B. Intensive Open-Ended Interviewing Techniques

Intensive open-ended interviewing techniques were used to gather every possible detail on subjects' mental processes in answering the technical questions. To gather this information, the subjects were instructed to think out loud as they solved the problems. This general technique is used in educational psychology to track the individual's comprehension on some lesson. In particular, attention was paid to the individual's cues (information from the question that is used in solving the problem), assumptions, and algorithms. Algorithms are defined as the heuristics used or mental operations performed by the subjects.

A simple ethnographic technique was used to obtain in-depth information, in the subject's words, on any bit of information that he mentioned. This technique is used in anthropological and other settings where there is danger of biasing the subject's account. This technique involves restating the subject's words into a question. For example, on beginning the randomness question (3), one subject said "Exactly five in a cell would be suspicious." He was asked, "Suspicious?" He replied, "Nonrandom. It wouldn't look random."

The subject's algorithms were identified using this technique and the more general technique of having the subjects think aloud. For example, the same subject examined the frequency table after giving his definition of "suspicious." He said, "These are clustered around five." Then he looked at the series and said "The four 8's in a row don't look random, but I've seen this before in longer runs, so it may be random. I don't see any other obvious patterns." He was asked, "Obvious patterns?" He replied, "Yes, if you can find runs of numbers like the 8's or alternating high and low numbers, it may not be random." His algorithm involves testing for randomness in the given series of numbers by evaluating the frequencies of the digits (provided in the table) and by looking for patterns in the series.

The assumptions that individuals unconsciously made were also investigated in this same manner of questioning. For example, when the same subject marked the answer to question 3, he said, "It's not too bad, so it may be random." He was asked, "Are you saying it may be random because you haven't found anything too suspicious?" He replied "Yes, I'm thinking that it's random unless it's obviously not." His assumption here is that the numbers are random unless proven otherwise by some testing procedures.

C. Data-Driven Analysis

From the research design, it was known that the interview data would be a mixture of quantitative data from background questions and from answers to the technical questions and of highly qualitative data from the in-depth interviewing process. It was also known that the effects being sought, that is, the sources of correlation, were difficult to find. This problem was compounded by only having a sample of 18 subjects available. The general philosophy of the proposed analysis was to keep data analyses simple and to let the data reveal their own secrets.

Prior to the interviewing, we felt that such a simple approach could be taken by careful examination of the data without imposing any models on the data. Hypotheses were formulated from the results of this data examination. These hypotheses would then be formally tested using very general statistical tests capable of handling the mixture of data. Analysis procedures are further described in Section IV.

D. Other Aspects of the Research Design

To test the remaining possible sources, the following design features were used. Technical questions were deliberately varied in their formulation. Appendix I contains the questions as seen by the subjects. Questions were included which had one verifiable answer (1, 2, 4), no verifiable answer (3), or more than one acceptable (verifiable) answer (5). Questions were constructed using a familiar problem formulation (1, 3, 5), or using a reversed formulation (2, 4). In the reversed formulation the information that is customarily given in the question was being asked as the answer, and the information that is customarily the answer was provided as the question.

Questions were asked on the subject's training and experience. They were asked about the focus of their schooling and work experience. These questions were asked prior to the technical questions. Appendix II contains the list of background questions. To further investigate work experience as a source of correlation, three subpopulations of the original 18 statisticians were interviewed: current members of the Statistics Group at Los Alamos National Laboratory, Los Alamos, NM, new members, and exmembers who had left the group but who maintained contact.

After the intensive interview on problem solving for each technical question, the subjects were asked how long ago they had worked on a similar problem. They were also asked if they had worked on that problem as a student or as an employee. The first of these questions was designed to test the hypothesis that the recency of exposure to the problem might affect the subject's solution. The second question was included because if common experience or schools of thought were identified as sources of correlation, information would be needed on whether the subject was using an approach he had learned at school or at work. This information plus that earlier obtained on the subject's professional background would allow individuals who had last worked on such a problem at school to be compared by school and those who had last worked on it at work by employer.

The duration of each interview was recorded. It was thought that the amount of time spent might relate to problem-solving style or accuracy because the more time-consuming methods of problem solving have this effect [6].

IV. Analysis of Results

A. Sources of Interexpert Correlation

The questionnaire design described in the previous section called for requesting and recording large amounts of information on each subject. The background questions alone provided a large multidimensional analytical problem. The first analysis step was to determine if any of the information from the background variables was commonly shared. If so, the original set of variables could be trimmed down to a set of independent variables that still retained a large portion (90 percent) of the original information. Factor analysis was used to investigate the structure of the background variables. By examining the factor loadings from a factor analysis and by using principles from judgment theory and common sense, a resulting set of independent variables was formed. The trimmed set included variables describing the number of years in the Statistics Group at Los Alamos National Laboratory, the number of years of statistical training, the percentage of current work involving walk-in types of problems, whether the respondent's degrees were in statistics, and whether the respondent was currently a member of the Statistics Group at Los Alamos.

To find sources of correlation among experts, it is necessary to find which variables (characteristics) were common to those subjects giving the same answers to the technical questions. Examining one technical question at a time, the trimmed set of background variables were analyzed with graphical and tabulation exploratory analysis to see if answers clustered according to the values of these variables. At this time, no strong evidence emerged for the background characteristics to be sources of correlation.

Therefore, the remainder of the information was examined in this exploratory analysis to look for potential sources. This information included the recency information and characteristics relating to the information gathered on the subjects' problem-solving processes. These characteristics were formulated as simple statements reflecting the assumptions and algorithms used by the subjects. For example, on question 2 five subjects used the assumption that the given correlation coefficient was large and used an algorithm describing a well-known correlation/sample size relationship. In characterizing the problem-solving steps, it was noted that for each question all the subjects used only a limited number of different assumptions (two or three) and a limited set of algorithms (one to three) and only a limited set of combinations of algorithms and assumptions (three to seven). The problem-solving steps could, therefore, be represented by these assumption–algorithm combinations. These combinations were called problem-solving pathways.

Exploratory analysis of these pathways produced the first evidence of sources of correlation. The experts' answers were indeed clustering according to their solution pathways. Table I illustrates what was seen by the authors when the assumptions and algorithms were tabled with the answers for question 2. These two variables were among magnitudes of qualitative data gathered during the interviews. Such data, in their entirety, could not be feasibly condensed to sets of tables, so this one example is given.

Such exploratory analyses of one variable at a time should not be interpreted as the final step. However, these analyses served to establish hypotheses on potential sources of correlation which were then tested using other statistical methods. For example, general linear models were used to statistically verify the relationships hypothesized between the pathways and background information and the answers. These models allow for the simultaneous examination of many variables to determine if their given values correspond to particular values of the answers.

The only difficulty in the construction of these models was in formulating usable variables to represent the pathways. This was finally done by constructing dummy variables (0 or 1) for dichotomous characteristics and rank variables (0,1,2) for such ordered charcteristics as "missing," "small," or "large." Separate variables were formed for each assumption and algorithm in this manner. The remaining information, such as background, recency, and duration of the interview, posed no modeling difficulties. The assumptions required for using linear models were monitored (i.e., normally distributed errors, linearity).

The SAS statistical analysis computer package has a procedure called GLM for analyzing general linear models. Since most answers were either strictly numeric or could be considered as ordinal ranks, application of these models was appropriate. Because question 3 responses were dichotomous, categorical analysis procedures (such as the SAS procedure FUNCAT) were also used. The sole purpose of these analytic procedures was to test the hypotheses concerning the potential sources of correlation. Models initially included large numbers of potential source variables. Using variate reduction procedures (stepwise regression), variables that did not appear promising as sources were analytically eliminated.

The results of the linear models on the final set of independent variables strongly indicated that many of the pathway characteristic variables provided good sources of correlation to the answers. These pathway variables alone accounted for as much as 78 percent of the total model variation. By contrast, the background

TABLE I
CLUSTERING OF ANSWERS BY PATHWAY FOR QUESTION 2

	Path 1	Path 2	Path 3	Path 4
Assumptions That the correlation is...	large	not considered	small	small
Algorithms As the sample increases, the correlation...	does not have to be large	is not considered	has to be large	does not have to be large
Answers				
	4	12	5 5	20
	8 8 8	15 15		23
	9	17		25
		18		27 27
				30

TABLE II
SUMMARY OF LINEAR MODELS' RESULTS

Variables	Emerges as a Source of Correlation[a]
Subjects' pathways (Assumptions) (Algorithms)	six out of six[b] questions
Subjects' education	
Degree in statistics	none
Work experience	
Statistical group member	none
Years in statistical group	none
Percentage of current work on these type of problems	one out of six questions
Recency of experience	one out of six questions
Duration of interview	one out of six questions

[a] A significance level of five percent was used.
[b] Question 3 had two parts, making a total of six.

factors, duration of the interview, and recency of experience were not significant sources of correlation. A summary of the variables which were significant sources is given in Table II. Details of these results are not presented because the purpose of the models was to test the significance of the hypothesized variables as sources of correlation and not to use the models in a predictive capacity. In each question the dominant source of correlation of answers is one or more pathway variables. While some other variables occasionally enter the models as sources, they are not the primary variables and are not present for all questions.

In examining the results of the linear models, a new hypothesis was tested to attempt an explanation of why subjects chose certain pathways. The hypothesis was that aspects of subjects' professional backgrounds led them to select particular paths, and thus that their backgrounds might indirectly influence the answers. However, no supportive evidence for this hypothesis was found from the linear models analysis.

B. Accuracy Issues

Because the problem-solving pathway variables emerged as the major sources of correlation, and because correct answers were available for most questions, an interesting question could be posed concerning the relationship of the pathways to the correct answers. That is, if similar pathways led to the same answers, which pathway(s) led to the correct answer?

In investigating this issue we found that the pathways leading to the correct answers had certain features in common and could be classified by the types of algorithms and assumptions used. To illustrate, in Table I the path labeled 1 is described with a simple and memorable relationship for the algorithm and with an assumption that correctly describes the problem. Subjects using this pathway arrived at the answers near the correct answer of 7. It is interesting to note the importance of the assumptions and algorithms used relative to the correct answers. Any conclusions from these results are probably artifacts of this study, and it is doubtful that they could be generalized. Certainly, if answers are verifiable, it would be advantageous to investigate the problem-solving features that led subjects to the correct answer.

Accuracy might also be related to the question complexity. It was hypothesized that the design characteristics of the questions might contribute to their complexity. These characteristics were the following:

1) whether one suitable algorithm was available for solving the problem or not;
2) whether the problem did or did not require the making of a particular assumption;
3) whether the post-survey answer variation was small or large;
4) whether the question had single or multiple correct answers;
5) whether the question had a familiar or reversed formulation.

A complexity index was formed for each of the questions by rating them on these five characteristics. These ratings were summed so that questions with more complex features had a higher sum. The percent of correct answers was used as the dependent variable, and the sum of the number of complex variables was used as the independent variable in a general linear model. As might be expected, the model indicated that more correct answers were given for the simpler questions. This result is consistent with the studies cited in Section II.

C. Aggregation of Expert Estimates

In using expert opinion data the decisionmaker needs a way of combining all the expert information for use by the decisionmaker. These opinions could be used to form a likelihood distribution to combine with the decisionmaker's prior distribution in a Bayesian context as advocated by Morris [15]. Alternatively, the opinions could be combined into a single estimator for use by the decisionmaker as a step in forming a likelihood or for use in the classical context. In focusing on the single estimator approach, a single estimate is usually thought of as representing a central measure (e.g., mean or median) of the distribution of all possible estimates. Correspondingly, an estimate of the dispersion or range of the values (e.g., a variance or confidence interval) is also desired to reflect the variability of the estimates. Previous studies (see [9]) have used many elaborate estimators to combine the individual estimates. They propose that the estimator that consistently provides the best coverage of the correct answer is the median. For these reasons, results will be presented here using the median, although the mean and geometric mean were also examined because of their popularity and for verification purposes.

Since pathway variables provided a key to the correlation of estimates, it was hypothesized that this information could be used in choosing an aggregation scheme. The three estimators of the entire sample for each question were computed and then compared to the estimators of subsets of the sample divided by the different pathways. For proper comparisons of the sample and subsample estimators, a measure of variability is required.

To obtain variation estimators for the mean, median, and geometric means, the distributions of these estimators for the sample and subsamples are needed. Assuming any distribution for these estimators would be pure speculation. Because the subsamples are so small and the data in these are skewed, even evoking the Central Limit Theorem (which states that the distribution of the mean is asymptotically normal) would be hazardous. In cases where the distributions are unknown, empirical distributions can be used to obtain estimates of the unknown parameters such as the variance and percentiles of estimators (in this case, the mean, median, and geometric mean). These empirical distributions are formed from the data through the use of simulation techniques.

One method of forming these empirical distributions is the bootstrap [16]. The bootstrap technique uses the data in a simulation and sampling scheme without making any assumptions about the distribution of the data. It is especially helpful for small samples when the analyst is hesitant on using asymptotic properties to determine variances of estimators. It is also used in applications where the data have large uncertainties (fuzzy data) because the simulation will preserve and reflect those uncertainties. In cases where distributions require verification, the bootstrap is used as a cross-validation technique [16].

A single bootstrap sample is formed by randomly selecting values from the original data. After each selection, that datum is replaced back into the original data set. A new sample is formed with the same size as the original data set. By forming many such new samples (e.g., 1000) and calculating the estimator of interest (e.g., the median), a distribution of that estimator is formed. This distribution provides estimates of the variance of the estimator as well as confidence limits from the distribution percentiles.

Bootstrap distributions for the mean, median, and geometric mean for questions with quantitative answers were formed using 1) all the data, and 2) the data from the subjects using the different pathways. Figs. 1 and 2 show the resulting 90-percent confidence limits of the medians of all the data and one pathway for questions 2 and 4, respectively. The pathway shown in Fig. 1 is the pathway 1 from Table I, and the pathway shown in Fig. 2 is a pathway that had similar characteristics to pathway 1. Using the correct answers as a reference value, the coverages of the 90-percent interval estimates of the medians for the entire sample can be compared to the coverage for the pathway subsample median. In both figures the coverages are quite different for the entire sample (all data) and for a specific pathway subsample. This difference is the result of the correlation or dependence of the experts according to the pathway used. In Fig. 1 the pathway subsample covers the correct answer with a narrow interval, while the all-data sample does not cover the answer even with its much wider interval. Similar coverage results were found using the mean and geometric mean. These 90-percent intervals for the estimators were not surprising in view of the nature of the variability in the original data, and bootstrapping cannot increase the variation beyond what is present in the original data. The nature of the all-data sample and subsample variability can be seen in Table I for question 2. The 90-percent intervals are consistent with the variations indicated there.

The particular pathways illustrated in Figs. 1 and 2 were chosen by the authors acting as decisionmakers. The choice was based on our examination of the various pathways found in the data. The pathways were chosen which used the assumptions that coincided with one we had in mind when formulating the question. A decisionmaker could use the information from the pathways to choose the estimate of a highly consistent group that exhibited some desirable feature, such as using an important assumption.

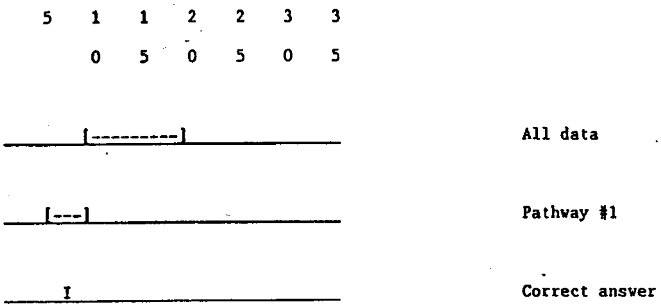

Fig. 1. Bootstrap 90-percent intervals for medians for question 2. Fifth and 95th percentiles for bootstrapped medians.

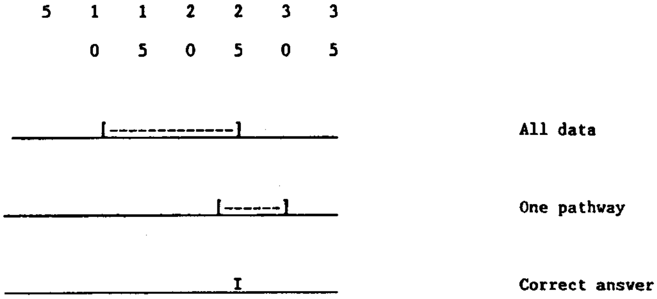

Fig. 2. Bootstrap 90-percent intervals for medians for question 4. Fifth and 95th percentiles for bootstrapped medians.

Conversely, a decisionmaker could use the pathway information (and the dependence structure reflected by it) to obtain a more diverse estimator. Since the pathways were found to be significant sources of correlation, samples constructed across the pathway groupings would be "pseudo-uncorrelated" samples and could be analyzed without violating the independence assumption. When such samples were formed from the data on questions 2 and 4, the medians of these pseudo-uncorrelated samples provided better coverage of the correct answer than did the medians of all the data (which are correlated data). However, the bootstrapped confidence intervals were much wider. This is a direct result from distribution theory which states that correlated samples have smaller variances than do uncorrelated samples. The creation of these pseudo-uncorrelated samples is another illustration of the usefulness of identifying the sources of correlation.

V. Conclusion

Expert estimates were found to correlate to the experts' problem-solving techniques and not to any features of their professional backgrounds. The paths that experts used to reach solutions included algorithms and assumptions. Assumptions and algorithms were found to correlate to the answers because they are integral parts of the path. In addition, assumptions were found to correlate to the accuracy of the answers: correct assumptions lead to correct answers. While this may seem an obvious result, it serves to emphasize the importance of specifying or eliciting assumptions [8].

From these conclusions, a few recommendations can be offered in eliciting expert opinion. It is recommended that the subject's problem-solving processes be elicited and recorded. This information could be used in evaluating the correlation structure, an evaluation that often needs to be performed in selecting and aggregating expert judgments. For example, if the estimates were correlated, the decisionmaker would need to use one of the aggregation schemes for dependent data. These schemes require knowing the structure of the correlation [10], [11].

In situations where the experts' estimates are being elicited in an interactive group situation, correlation between expert and decisionmaker estimates are likely to occur. This dependence has been called the follow-the-leader or group-think effect [14]. Obtaining participants' problem-solving data is more difficult in this setting than with isolated experts because of the group dynamics [12]. However, the problem-solving information would be valuable in determining the structure of correlation and determining which estimates to include and how to aggregate them. How the problem-solving features might be affected by the group situation is a complicated area and worthy of future study.

In the case where there is only one expert, the decisionmaker would benefit from having the problem-solving data because he may need to combine the expert's estimate with his own. Aggregating estimates in this situation, using a Bayesian framework, requires that the decisionmaker know enough about the expert's estimate to answer two questions: 1) how much his estimate overlaps with that of the expert [17], [18], and 2) the accuracy, or calibration, of the expert [18]–[20]. Dealing with dependence and calibration are considered tricky problems [21]. Information on what data the expert considered and how he solved the problem will allow the decisionmaker, in this scenario, to best assess the overlap between his prior estimate and the expert's data. Also, if the decisionmaker is trying to calibrate the expert to weight his estimate in the aggregation formula, he can look at the expert's problem solving. In this way, he can check such things as whether the expert followed the directions, agreed upon assumptions, and so on. He can perhaps identify careless errors, such as in the expert's substitution of numbers in an equation, and consider this information in weighting the expert's judgment.

In general, it is recommended that the questions asked of the experts be constructed to include as much specific information as possible to aid the subject in solving the problem. In particular, any information that defines the question, such as assumptions to be made or definitions to be used, should be included. In addition, any aids needed in solving the question, such as equations, conversions, or states, should be provided to all the experts. If this information is provided, the experts are less likely to make errors resulting from incorrect recall or from using an inappropriate assumption.

It is also recommended that there be other field studies into sources of correlation. It would be helpful to look at sources/structure of correlation in different settings such as in more complex question environments and group elicitation situations. We have recently completed a study examining the aforementioned results in a more complex question environment [22].

Appendix I

Question 1: Chi-Square Goodness of Fit

The following observed values (O) are counts of samples of maize plants from a Mendelian experiment. The theoretical ratios and corresponding expected values (E) are from Mendelian inheritance laws:

Sample	Ratio	Expected	Observed	$O-E$	$(O-E)^2/E$
Green	9	738	752	14	0.27
Golden	3	246	231	−15	0.91
Green-stripe	3	246	230	−16	1.04
Golden-green	1	82	99	17	3.52
		1312	1312		5.74
		degrees of freedom = 3			

The above chi-square test statistic is significant at the following values:

a) one percent (meaning at or less than 0.01);
b) five percent (meaning greater than 0.01 but less than or equal to 0.05);
c) ten percent (meaning greater than 0.05 but less than or equal to 0.10);
d) ten percent (meaning greater than 0.10).

Question 2: Pearson Sample Correlation Coefficient

A sample correlation coefficient between two measurements of geologic core samples is 0.70. Draw a line at or between the following sample sizes where this value of r is significant at the five-percent level for a one-tailed test:

$n =$ 1
5
10
15
20
25
30.

Question 3: Randomness

The following is a set of 50 numbers (grouped by fives):

06318 37403 49927 57715 50423
67372 63116 48888 21505 80182.

The following is the frequency table for the ten digits:

x	f
0	5
1	6
2	5
3	6
4	4
5	5
6	4
7	6
8	7
9	2

Part A: Is the above set of digits random? (Use a five-percent significance level):

YES or NO.

Part B: Is the following set random by the same criterion?

69301 22875 49382 43635 01312
13898 98878 76421 90121 01564.

x	f
0	4
1	8
2	6
3	6
4	4
5	3
6	4
7	3
8	7
9	5

YES or NO.

Question 4: Central Limit Theorem

The histogram of Fig. 3 is of an original population from which samples are to be taken. The histogram of Fig. 4 is of means of samples (all the same size n) from the original population above. With the Central Limit Theorem as a guide, what is the value of the sample size n? Choose from the list below by

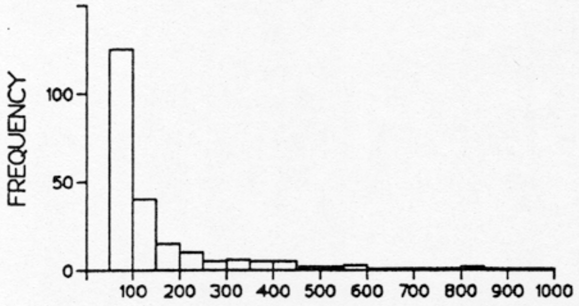

Fig. 3.

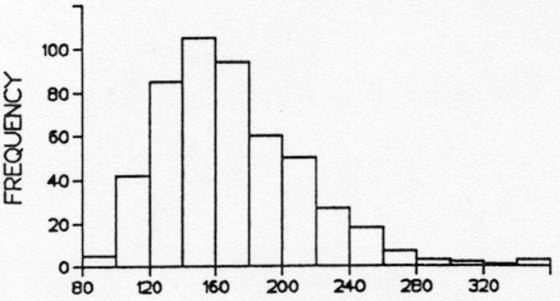

Fig. 4.

drawing a line at or between the given values:

$n =$ 5
10
15
20
25
30.

Question 5: Analysis of Variance—Multiple Comparisons Tests

The following sets of test scores were collected for three different aptitude tests from a randomly selected set of students:

Test	Scores				Sum	Mean
AA	80	90	87	83	340	85
BB	70	82	78	74	304	76
CC	63	76	70	71	280	70

The following table shows the results of a one-way AOV:

Source	Degrees of Freedom	Sums of Squares	Mean Squares	F Test Statistic
Test	2	456	228	9.16
Error	9	224	24.9	
Total	11	680		

If the desired alpha level is five percent, which option reflects the differences (if any) existing between the test means for tests AA, BB, and CC? We have the following:

a) $\underline{85\ 76\ 70}$ (all means are the same);
b) 85 76 70 (all means differ from each other);
c) $85\ \underline{76\ 70}$ (85 is larger than 76 and 70; 76 and 70 are the same);
d) $\underline{85\ 76}\ 70$ (70 is smaller than 85 and 76; 85 and 76 are the same);
e) $\underline{85\ 76}\ 70$ (85 and 76 are the same; 76 and 70 are the same; but 85 is larger than 70).
$\underline{76\ 70}$

Appendix II
Professional Background Questions

Name____________________

1) Where do you now work? Institution____________________
Group/Dept.____________________

2) For how many years? Years____________________

3) (If not in S-1, ask) Have you ever worked in or consulted for S-1 at Los Alamos National Lab? Yes________ No________

4) (If yes to 3, ask) For how many years? Years____________________

5) (If yes to 3, ask) How many years have elapsed since then? Years____________________

6) About how many years of statistical experience—from school or work—have you had? Years____________________

7) Have you done any consulting—helping others solve real problems? Yes________ No________

8) (If yes to 7, ask) For how many years? Years____________________

9) About what percentage of your current work could be described as being classic walk-in consulting problems? Years____________________
How would you describe the remaining work?____________________
____

10) What degrees do you hold in statistics? B.S.___ M.S.___ Ph.D.___

11) At what schools did you receive degrees in statistics?
Undergraduate__________ Theoretical___ Applied___
Graduate__________ Theoretical___ Applied___
Graduate__________ Theoretical___ Applied___

12) (For each school, ask) Was the statistics department considered theoretical or applied?

13) Which of the following do you enjoy most about your profession?
The mathematical aspects ___
The broad-based applications ___
The challenging problems ___
The interaction with people ___
Other____________________

After the subject completes the technical questions ask the following:

14) Did you think that these questions were
Too easy _____
About right _____
Too difficult _____
Unclear _____

15) Do you feel that these statistics questions tapped your expertise?
Not at all _____
Slightly _____
Half the time _____
Most the time _____
All the time _____

16) From what areas of statistics would you have drawn questions to better tap your areas of expertise?
1.____________________
2.____________________
3.____________________
4.____________________

Acknowledgment

The information provided by the individuals in the Statistics Group of Los Alamos National Laboratory is gratefully acknowledged.

References

[1] N. C. Dalkey, "Toward a theory of group estimation," in *The Delphi Method*, H. Linston and M. Turoff, Eds. Reading, MA: Addison-Wesley, 1975.

[2] G. B. Baecher, "Correlations among experts' opinions," Mass. Inst. Technol., Boston, MA, unpublished manuscript, 1979.

[3] R. Hogarth, *Judgment and Choice: The Psychology of Decisions*. Chicago: Wiley-Interscience, 1980.

[4] A. Tversky and D. Kahneman, "Framing of decisions and the psychology of choice," *Science*, vol. 211, pp. 453–458, Jan. 30, 1981.

[5] M. K. Comer, D. A. Seaver, W. G. Stillwell, and C. D. Gaddy, "Generating human reliability estimates using expert judgments," Sandia Nat. Lab., Albuquerque, NM, Tech. Rep. NUREG/CR-3688, SAND84-7115, 1984.

[6] B. Hayes-Roth, "Estimation of time requirements during planning: Interactions between motivation and cognition," Rand Corp., Santa Monica, CA, Tech. Rep. N-1581-ONR, 1980.

[7] M. Matz, "Towards a process model for high school algebra errors," in *Intelligent Tutoring Systems*, D. Sleeman and J. S. Brown, Eds. New York: Academic, 1982.

[8] W. Ascher, *Forecasting: An Appraisal for Policymakers and Planners*. Baltimore, MD: John Hopkins Univ. Press, 1978.

[9] H. F. Martz, M. C. Bryson, and R. A. Waller, "Eliciting and aggregating subjective judgements—Some experimental results," in *Proc. 10th Ann. SAS Users Group Int. Conf.*, SAS Institute Inc., Cary, NC, 1985.

[10] R. L. Winkler, "Combining probability distributions from dependent information sources," *Management Sci.*, vol. 27, no. 4, pp. 987–997, Apr. 1981.

[11] D. V. Lindley and N. D. Singpurwalla, "Reliability and fault tree analysis using expert opinions," George Washington Univ., Washington, DC, Tech. Rep. GWU/IRRA/TR-84/10, 1984.

[12] M. A. Meyer, A. T. Peaslee, Jr., and J. M. Booker, "Group consensus method and results," Los Alamos Nat. Lab., Los Alamos, NM, Tech. Rep. LA-9584-MS, 1982.

[13] P. A. Seaver, "Assessments of group preferences and group uncertainty for decision making," Social Science Res. Inst., Univ. of Southern California, Los Angeles, CA, 1976.

[14] I. C. Janis, *Victims of Group Think: A Psychological Study of Foreign Policy Decisions and Fiascos*. Boston, MA: Houghton Mifflin, 1972.

[15] P. A. Morris, "Combining expert judgments: A Bayesian approach," *Management Sci.*, vol. 23, no. 7, pp. 679–693, Mar. 1977.

[16] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, Jan. 1979.

[17] R. T. Clemen, "Calibration and the aggregation of probabilities," *Management Sci.*, vol. 32, no. 3, pp. 312–314, Mar. 1986.

[18] M. J. Schervish, "Comments on some axioms for combining expert judgements," *Management Sci.*, vol. 32, pp. 306–312, 1986.

[19] S. French, "Calibration and the expert problem," *Management Sci.*, vol. 32, no. 3, pp. 315–321, Mar. 1986.

[20] P. A. Morris, "Observations on expert aggregation," *Management Sci.*, vol. 32, no. 3, pp. 321–328, Mar. 1986.

[21] R. L. Winkler, "Expert resolution," *Management Sci.*, vol. 32, no. 3, pp. 298–303, Mar. 1986.

[22] M. A. Meyer and J. M. Booker, "Sources of correlation between experts: Empirical results from two extremes," Los Alamos Nat. Lab., Los Alamos, NM, Tech. Rep. NUREG/CR-4814, LA-10918-MS, 1986.

0018-9472/88/0100-0142$01.00 ©1988 IEEE