Determining Differences in Expert Judgment: Implications for Knowledge Acquisition and Validation*

Daniel E. O'Leary
Graduate School of Business, University of Southern California, Los Angeles, CA 90089-1421

ABSTRACT

In knowledge acquisition, it is often desirable to aggregate the judgments of multiple experts into a single system. In some cases this takes the form of averaging the judgments of those experts. In these situations it is desirable to determine if the experts have different views of the world before their individual judgments are aggregated. In validation, multiple experts often are employed to compare the performance of expert systems and other human actors. Often those judgments are then averaged to establish performance quality of the expert system. An important part of the comparison process should be determining if the experts have a similar view of the world. If the experts do not have similar views, their evaluations of performance may differ, resulting in a meaningless average performance measure. Alternatively, if all the validating experts do have similar views of the world then the validation process may result in paradigm myopia.

Subject Areas: Decision Support Systems, Expert Systems, Group Decision Processes, and Verification Theory.

INTRODUCTION

Sometimes in the knowledge acquisition process, developers of expert systems have access to the judgment of multiple experts. In those cases it is likely that the developer may have to aggregate across the experts to develop a single system, possibly averaging expertise. However, depending on the purpose of the system, the developer probably should not aggregate the judgments of experts from different camps (i.e., experts from different firms or experts with distinctly different views of the world). For example, with a political expert system it is difficult to envision who would be interested in a system that averaged the advice of a conservative and a liberal. However, expert incompatibility is so easily established in few cases.

Similarly, in the validation process a Turing test [11] can be implemented to determine the overall quality of the system by using multiple experts to evaluate the comparative quality of an expert system and other human actors. In such evaluative comparisons, it can be important to know if those multiple experts have similar views of the world. Otherwise, the development of an average score to compare the system to other human actors could be a meaningless integration of

*The author wishes to acknowledge the helpful comments of the anonymous referees and associate editor on earlier versions of this paper.

judgments of apples and oranges. Alternatively, if all the validation comes from experts from the same school there may be paradigm myopia. In that situation, assumptions and important aspects missing from the system may be assumed away by expert evaluators steeped in the same paradigm.

The purpose of this paper is to investigate the determination of whether multiple experts (in either a knowledge acquisition or validation situation) have similar views of the world. As a result, this paper also provides a comparison of some statistical approaches. This paper accomplishes those purposes in the following manner. The next section provides further discussion on the importance of determining whether experts view the world from a similar perspective. Two case studies are presented which illustrate the approach used throughout the remainder of the paper. Then the different statistical approaches used to evaluate the existence of similarity of views of the multiple experts are discussed. Next the use of traditional and computer intensive statistics, using the case study data, to assess the existence of differences between experts are investigated. The following section extends the approach to investigating measurement of the existence of differences between groups of experts. The last section briefly summarizes the paper and investigates some potential future research issues.

DIFFERENCES IN EXPERT JUDGMENTS

As noted earlier the determination of differences in judgments among multiple experts is important in both knowledge acquisition and validation.

Knowledge Acquisition

The development of a system using multiple experts can take at least two approaches. First, each module of the system may employ a single different expert yet each of those modules would be linked to other modules in some cascading manner. Thus, one expert's judgment would cascade into another's judgment, and so on. It could be important to have consistent expertise modeled in each of the modules, or else decisions in one module might contradict decisions in another. Second, the system may use some sort of average expertise of multiple experts in each module. Such averaging takes place anytime knowledge from multiple sources is summarized into a single representation. In that case the average judgment may be built into, for example, the weights on rules. In order for that average to represent some consistent set of experts, it is important to determine if those experts are from the same camp.

Further, the examination of the similarity of judgment might be accomplished either by pairwise comparison of individual experts or pairwise comparison of groups of experts. These comparisons would be made if there was some basis to expect differences among individual experts or if there was some reason to assume that the experts from one group, department, or firm were different than the experts from some other group, department, or firm. These two sets of categories of analysis are summarized in Table 1.

There are four possible outcomes in the scenario developed in Table 1. The two sets (1,3) and (2,4) each require different minimal sets of information for the comparison to be made. First, it is assumed in cases 2 and 4 that there is sufficient

Table 1: System use of different experts.

Comparison Basis	Interlocking Expertise	Average Expertise
One-to-One	1	2
Group-to-Group	3	4

information to compare the two experts or sets of experts with each other for an entire set of decisions $D=(d_1, \ldots, d_n)$ where those decisions d_i, could be, for example, estimates of a weight on a rule. Second, for cases 1 and 3, to compare two sets of decisions, $D_1=(d_1, \ldots, d_{n1})$ and $D_2=(d_{n1+1}, \ldots, d_n)$, respectively, pairs of experts or groups of experts are required.

These different sets of information suggest that different approaches be used to analyze the data. If for cases 1 and 3 there is information for a complete set of decisions, then an analysis, similar to that for cases 2 and 4, may be used. However, such availability of information is unlikely, since the rationale for developing modules with different experts is often the limited time of experts to contribute to the project or because different levels of expertise are needed with each module or portion of the decision.

The focus of this paper is on cases 2 and 4. In either of those cases the similarity of experts' judgments across an entire set of decisions is the basis of analysis.

Validation

Typically implementations of the Turing test employ experts to rate the performance of expert systems and other comparative human actors. The system and these other human actors are given test problems and then rated for the quality of their performance. The rating is likely to be on a scale of 0 to 10 where 0 is the lowest quality rating and 10 is the highest, or 0 and 1, where 0 is unacceptable and 1 is acceptable. Then those ratings often are averaged together to provide a measure of the overall quality.

The development of such an overall average measure of quality could lose its meaning if the experts have different perspectives. In that case, combining the evaluations would be analogous to adding disparate commodities (apples and oranges). Alternatively, if all the experts in the validation process are from the same school then that may indicate that important aspects of the domain will be ignored or assumed away (paradigm myopia). Thus, in either case, it is important to determine the experts' view of the particular domain.

SYSTEM CASE STUDIES

This paper analyzes two case studies in the development and validation of expert systems. Although this paper concerns itself only with these systems, the methods it discusses are applicable to other systems. The discussion of the examples involves relatively small samples. However, it is not unusual in practice for larger samples of experts and experts' decisions to be available.

Aggregating the Judgment of Multiple Auditors

Some expert systems aggregate the judgments of multiple experts into a single judgment. AUDITOR, an expert system, (Dungan [1, p. 1] and Dungan and Chandler [2]) "presents expert advice in the form of an estimate of the probability that a given account balance will prove uncollectable." The weights on the rules in the system were based, in part, on the expertise of four auditor experts. The assessments of those four experts are summarized in Table 2.

If the experts are disparate in their estimates then it may not be reasonable to aggregate their judgments into a single estimate of the weights on a rule. For example, experts from one firm may respond differently to different criteria or different experts may have different systems of belief. Thus, it may be appropriate to build the model around only a subset of the experts. It may be more efficient to use the judgments of experts from a single camp or firm to build the estimates.

Since the assessments of the weights are treated as numeric in the development of the system by Dungan [1], one approach is to give each interpretation an integer value (e.g., strong=4, moderate=3, etc.). Using those estimates, different test statistics can be developed and tested for their statistical significance.

Validation Using Multiple Experts

Hochman and Pearson [5] [6] developed an expert system, X-Breed, designed to choose between eight different breeding strategies and to recommend the bull breed from a list of ten types of breeds. The system was evaluated using a comparison of the system to ten human evaluators, including students, advisory officers, case providers, and research officers.

For nine different problems eleven responses were rated as either acceptable or not acceptable solutions by each of four independent experts. (Summaries of the nine different cases are included in Table 3.) Then the average of the number of acceptable solutions was used to rank the system and the human actors. The study used the same approach as in Yu, Fagan, Wraith, Clancey, Scott, Hannigan, Blum, Buchanan, and Cohen [12]. However, [5] and [6] provide more detail to allow for further investigation.

The Unit of Analysis

In the case of aggregating the judgments of multiple expert auditors, it is assumed that the entire set of rules is an appropriate unit of analysis. The set of rules is treated as an aggregate; within that aggregate we can expect that experts are from one camp or another. If there is some reason to assume that different rules should be treated differently then our analysis would take that particular theory into account. It may be that instead of using all rules as the unit of analysis subsets of rules should be analyzed separately. However, throughout the remainder of the paper it is assumed that the rule-base should be analyzed as a whole.

Similarly, in the case of validation using multiple experts, it is assumed that the entire set of advisors is an appropriate unit of analysis throughout this paper. The set of advisors is treated in the aggregate. If there is some reason to evaluate some subset of the advisory officers differently, then that should be accounted for

Table 2: Relative strengths of the rules.

	Auditor				
Rule Name	A	B	C	D	Composite
Active	S	W	S	S	S
Bankrupt	S	S	S	S	S
Collected	S	S	S	S	S
Correspond	S	M	S	S	S
CreditMgr	W	M	M	M	M
Economic	W	S	M	S	M
Noncontact	M	M	M	M	M
NoResponse	W	W	W	M	W
NotPay	S	S	S	S	S
Problems	M	W	W	M	W-M
Rigorous	W	W	W	N	W
Workout	S	M	S	S	S
Writeoff	M	W	W	M	W-M

S = Strong
M = Moderate
W = Weak
N = Little or no effect

Source: Dungan [1, p. 95]

in the analysis. Since there was no information regarding the specific type of advisor in the information provided to the experts, there is no a priori reason that the experts would treat the responses from any advisor differently.

In general, this indicates that the unit of analysis (ultimately, the data set analyzed) is a critical issue. At some point in the process of statistically analyzing knowledge acquisition or validation data, there should be an investigation into the choice of the appropriate unit.

That choice is, to a large extent, a function of the specific application and outside the scope of this paper. However, if there is a theory or rationale that indicates that different subsets of the data should be treated separately or together, then the analysis should exploit the theory. In the case of the aggregation of weights on rules, the aggregate is assumed as the appropriate basis of analysis for a number of reasons: the rules form a cohesive subject group; the rules all deal with similar aspects of the same decision problem; the rules were all developed at the same time; and other similar reasons.

STATISTICAL BASIS OF ANALYSIS

The choice of a test statistic requires careful consideration. The mean value of one expert's evaluations probably does not provide a suitable basis for comparison of the mean value of another expert. Simply because the mean values of the expert judgments are the same, does not ensure that the way the experts or groups of experts assess the individual rules is the same. For example, one expert might

Table 3: Multiple expert evaluation of recommendations.

Advisor	Score Expert 1	Score Expert 2	Score Expert 3	Score Expert 4	Mean Score
X-breed(system)	9	9	7	9	8.50
Research Officer	9	9	7	9	8.50
Advisory Officer	9	7	6	7	7.25
Case Providers	8	7	6	7	7.00
Advisory Officer	7	7	5	8	6.75
Advisory Officer	7	8	5	6	6.75
Advisory Officer	5	8	6	6	6.25
Advisory Officer	7	6	5	7	6.25
Advisory Officer	5	6	4	6	5.25
Advisory Officer	5	2	4	6	4.25
Student	4	1	0	3	2.00

Note: Adapted from Hochman and Pearson [5].

assign two rules weights of 4 and 1, respectively, while another expert might assign those same two rules values of 1 and 4, respectively. Clearly, the average would be the same, yet the understanding of the rules for the process would be different for each of those experts.

Alternatively the correlation coefficient provides pairwise comparison of different components of experts' judgements simultaneously. Thus, the correlation between the assessments (either estimates of uncertainty factors or evaluations of quality) made by different experts is a critical concern. If the assessments are significantly correlated then we reject the notion that the evaluations come from different distributions. In that situation, the experts would be specified as coming from the same camp. The evaluation of the significance of the correlation coefficient can employ either traditional distributional approaches (e.g., a *t*-test) or computer intensive approaches.

Traditional Approach

One of the most common approaches to the investigation of similarity of correlation coefficients is a *t*-test. This approach was used on both case studies. However, this approach assumes an underlying distribution to make such significance assessments. In many situations the assumptions associated with those distributional assumptions may be incorrect or inappropriate.

Computer Intensive Statistics

An alternative approach to the determination of significance are computer intensive statistical methods. These methods are called computer intensive because the computer recomputes the test statistics using the original sample a large number of times (e.g., 100, 500, or 1000 times) to develop a distribution of that test statistic (e.g., mean or standard deviation) for that particular sample (e.g., Noreen [9]).

Traditional statistical approaches typically make distribution assumptions. Under the assumption of a distribution (e.g., a normal distribution) quite powerful tests have been developed. However, in those cases where no distributional assumption is made, the resulting tests often are lacking. For example, Chebyschev's inequality [7] can be used to obtain an upper bound on the probability of a random variable being more than a specified distance from the mean of the distribution. Unfortunately, such approaches often provide very loose estimates, and as a result are of little use.

Making distributional assumptions may be inappropriate and could lead to questionable results. Since expertise and the process of gathering expertise generally are not well understood, it typically is difficult to make a priori distributional assumptions about the data. For example, often experts employed to assess the quality of an expert system provide integer estimates of comparisons between systems and human experts. In addition, typically those estimates are bounded (e.g., between 0 and 100 in the case of percentages or between 1 and 9 in the case of survey scales). These are not typical assumptions made on data for continuous distributions. Accordingly, an alternative to distribution-based statistics is desired to investigate expertise.

Unlike traditional statistical methods, computer intensive statistics make few, if any, distributional assumptions. Yet these methods have been found to provide powerful results. Instead of distributional assumptions, computer intensive statistics use the sample data to develop empirical distributions. The argument behind the use of the sample is that the sample contains all that is known of the underlying distribution (e.g., Efron [4]). Thus, computer intensive statistics can fill an important void.

This paper discusses three different types of computer intensive tests: enumeration (Noreen [9]), randomization (Edgington [3] and Noreen [9]), and bootstrap resampling (Efron [4] and Noreen [9]). Enumeration is where each feasible case is enumerated to establish the distribution. Enumeration involves enumerating all feasible combinations of a specific sample then computing the test statistic in order to develop an exact distribution of that test statistic. Randomization involves shuffling or randomizing one variable relative to another, computing a test statistic for that random variable, using that test statistic to continue to build a distribution, and then randomizing once again. Such a process is done a large number of times, to develop an approximation of the exact distribution, which may be computationally infeasible to develop.

Randomization generally is used to investigate the null hypothesis that one random variable is distributed independently of another random variable [9]. Bootstrap resampling assumes that the sample population is representative of the actual population. As noted by Efron [4] since the sample contains all the information about the population, we proceed as if the sample is the total population. Bootstrap resampling is simulation analysis, under the assumption that the simulation generates its observations by successively resampling from the sample data with replacement.

RESULTS OF THE PAIRWISE COMPARISONS OF EXPERTS

Both traditional and computer intensive approaches were used to analyze the data from the two case studies.

Knowledge Acquisition Case

Assuming that the assessments by the experts can be summarized as a set of integers (strong=4, etc.), the correlations between each pair of experts are summarized in Table 4. Using a *t*-test, each pair is found to be significant at the .01 level (and better) except for the comparison of experts A and B, which is significant at the .05 level.

For the computer intensive approach, a distribution of correlations for each pair of experts was developed. The results were not as statistically significant as those using the *t*-test approach. Using this approach, the correlation between A and B is not significant at the .1 level. In addition, both the comparison B:C and B:D are less significant using the computer intensive approach. These findings suggest that expert B judges weights on the rules differently than expert A. Thus, it may be beneficial in the development of the expert system to aggregate the judgment of experts in the set B, C, and D or the set A, C, and D.

Validation Case

Pairwise analysis of each of the experts in the validation process resulted in statistically significant correlation coefficients using both *t*-test and computer intensive methods. Using a *t*-test approach, each of the correlations in Table 5 was significant at the .001 level. However, as with the other case study, the computer intensive approach yielded a number of correlations where the test statistics were not as significant. For example, the correlation for the pair 1:2 was between the 92nd and 93rd observations.

These results indicate that the human experts in the validation effort apparently come from the same school. Unfortunately, if this is the only comparison there is potential for paradigm myopia.

Comparison of Statistical Approaches

The two case studies presented here also illustrate some of the differences that can occur with the use of, for example, a *t*-distribution and a computer intensive approach. In the case 1 comparison of A:B, the *t*-distribution estimate of statistical significance was at the .05 level, yet the computer intensive approach estimated the statistical significance at the .25 level. These two statistical estimates suggest alternative actions: combine versus not combine estimates of probabilities.

In addition, in both studies, measures of statistical significance differed from pair to pair. For example, in case 2, the pair 1:3 had a correlation of .783 that was found above the 100th observation in a distribution of 100 correlations. Pair 3:4 had a correlation of .886 and was located below the 100th observation.

These findings suggest that both computer intensive and *t*-distribution approaches be used to evaluate the significance of the correlation between different experts. If both are significant at an appropriate level then the experts would be judged as coming from the same school. Otherwise, the experts may not be from the same camp.

Test for Goodness of Fit of a Normal Distribution

Prior to performing a *t*-test, the knowledge engineer may wish to perform a test to determine the goodness of fit of a normal distribution to, for example, the expert

Table 4: Results of randomizing correlations on relative strengths.[+]

Pair	Correlation	Location*
A:B	.339	Approximately the 75th observation
A:C	.794	Above 100th observation
A:D	.664	Above 100th observation
B:C	.662	At 97th observation
B:D	.598	At 97th observation
C:D	.771	Above 100th observation

[+]Assumes a distribution of 100 elements. Based on Table 2.
*Roughly the significance level is (100–Location)/100.

Table 5: Results of randomizing correlations on recommendations.[+]

Pair	Correlation	Location*
1:2	.723	Between 92nd and 93rd Observation
1:3	.783	Above 100th Observation
1:4	.827	At 100th Observation
2:4	.776	At 98th Observation
2:3	.882	Above 100th Observation
3:4	.886	Between 99th and 100th Observation

[+]Assumes a distribution of 100 elements. Based on Table 3.
*Roughly the significance level is (100–Location)/100.

recommendations such as those in Table 3. The Kolmogorov-Smirnov (K-S) test of normality, with both the mean and the standard deviation estimated, provides one approach.

The K-S test is applied to expert 3 and summarized in Table 6. It is found that the hypothesis, the responses from expert 3 are drawn from a normal distribution, can be rejected at the .05 level.

Table 6 is constructed and evaluated using the table of critical values constructed in Lilliefors [8]. The first column is the set of values in the data. The second column is the cumulative frequency f_1 of those values. The third column is the ratio (data value−mean)/(standard deviation). The fourth column is the amount of area under the normal curve for the ratio from column three, f_2. Finally, column five is the absolute value of f_1-f_2. If the maximum value in column five exceeds the appropriate critical value, then the hypotheses can be rejected. The critical value is derived from Lilliefors [8], given a sample size of ten and significance level of .05. In this case, the critical value is .258, while the maximum value is .278, and thus the hypothesis of normality can be rejected.

The K-S test can be extended to any continuous distribution. In addition, other distribution tests could be used. For example, the chi-square goodness of fit test might be used.

Table 6: Kolmogorov-Smirnov goodness of fit test distributional analysis of expert #3.*

Data Value	Cumulative Frequency Percentage	z Ratio	Normal Curve Area	Column 2 - Column 4 Absolute Value
0	.1000	-5.092	.0001	.0999
4	.3000	-1.388	.0828	.2172
5	.6000	-.462	.3218	.2782
6	.9000	.462	.6783	.2217
7	1.0000	1.388	.9182	.0818

*Data for expert #3 is contained in Table 3.

EXTENSIONS: DETERMINING THE EXISTENCE OF DIFFERENCES DUE TO FIRM EFFECT

The investigation of the existence of differences due to different experts being from different firms requires an alternative approach. The existence of expertise in different firms (departments, etc.) uses a construct of at least one expert per firm construct, so that a correlation between two individual experts may no longer be appropriate. Instead the concern would be whether the set of experts' judgments from two different firms comes from the same distribution. In this situation, we either assume, or test to determine, that the experts' judgments within each of the individual firms come from the same distribution. Such testing can be done as was outlined above.

In the case where we know that experts come from, say, two different departments or firms (or some other theory), the null hypothesis could be that the experts' ranking of the strength of the rules is independent of, say, the firm for which that expert works. (A similar scenario would be constructed for validation.) Assume a matrix of data $x_{i,j}$ for each expert j=1, ..., m, and strength given to rule i=1, ..., n. Assume that m_1 experts work for one firm and m_2 experts work for another firm (m_1+m_2=m).

Enumeration/Randomization

In order to measure the significance of the difference assigned by the experts in the two firms we can use either an exact or approximate approach. In the approximate approach, we shuffle the vectors $\mathbf{X}_j$=($x_{1,j}$,...,$x_{n,j}$), where $x_{i,j}$ is the value on the ith rule for the jth expert, into two groups, of m_1 and m_2 vectors. Then we compute the test statistic (e.g., average scores for each i) for each of the two groups. This gives two vectors of averages. The average of the sum of the absolute values of the differences between the two vectors forms the value for that shuffle. This would be done a large number of times (e.g., 1000) in order to establish an estimate of the distribution of such absolute differences. Then the original sum of the absolute value of the differences for the two groups would be compared to the derived distribution to determine if there was a significant difference between the two firms. The computer generated distribution would be used to measure the significance of the sum of absolute differences of the original data.

The exact approach is similar to the approximate approach, except that each possible shuffle would be done exactly once. If m is reasonably small, such an approach is feasible. In general, this analysis would use groups of the size suggested by the theory (e.g., all members of one firm would be one group).

The small number of experts in the case examples makes it difficult to generate a reasonable sample size, even with complete enumeration of all possible combinations. Distributions based on groups of experts of size 2 and 2, and groups of size 3 and 1, were developed. The resulting two distributions are summarized in Table 7. These distributions are presented for illustrative purposes only.

Bootstrapping

An alternative approach to developing a distribution is bootstrapping (Efron [4]). The following approach could be used in order to investigate the null hypothesis that there is no firm effect.

Because the number of experts from each firm may be different, that must be accounted for in the test developed. Divide the sample into the two different sets of experts' vectors (e.g., by different firms) so that there are m_1 in one set and m_2 in the other. Bootstrapping uses sampling with replacement from those two sets. Randomly choose an expert's vector from the first set, keep track of the contents of the vector, replace that vector back into the first set, then choose another from the first set. Do that procedure m_1 times. Given those m_1 vectors, compute an average for each i, resulting in a vector of averages. Use the same procedure to develop a vector of averages for the second set (instead sample m_2 times from those observations). Then compute the sum of the absolute values of the differences between the two different vectors of averages.

Do this procedure a large number of times in order to develop a distribution. The resulting distribution can be used to assess the significance of the difference between the vectors of averages based on using each of the m_1 and m_2 elements, respectively. Unfortunately, the sample systems have only four experts. Such small samples make it difficult to illustrate the technique.

SUMMARY AND FUTURE RESEARCH

This paper addressed the determination of whether multiple experts came from different camps. The issue was identified as important to both knowledge acquisition and validation. This section briefly summarizes the paper and discusses some extensions.

Summary

Two comparative approaches were investigated: the one-on-one comparison of the similarity of pairs of different experts and the comparison of different, a priori, (e.g., firm employees) groups of experts. Throughout, it was assumed that the basic unit of analysis was the complete set of data available from each of the cases.

In the situation of one-on-one comparison of experts, correlation was used to ascertain the similarity of expert judgments. The analysis of the first case study found that one pair of experts used to develop the weights on the rules (A:B) appear

Table 7: Enumeration of two sets of groups ranked according to sums of absolute differences.

Group X	Group Y	Sum of Absolute Differences in Rules*
Groups of 2 and 2		
A, B	C, D	4
A, C	B, D	5
A, D	C, B	6
Groups of 3 and 1		
A, B, D	C	4
A, B, C	D	6
B, C, D	A	6
A, C, D	B	7.333

*Sum of absolute differences computed by developing average for one group and comparing average to other group, based on numerical scores assigned to the values in Table 2 of S=4, M=3, W=2, and N=1.

to come from different camps. This could result in the development of weights that do not represent a consistent set of experts. The analysis of the second case study found that each of the evaluating experts came from the same camp. Potentially, this could limit the quality of the validation effort because of paradigm myopia.

A *t*-test and randomization were used to provide statistical evidence. It was found, in some situations, that the two approaches had critical differences between the extent of statistical significance associated with the test statistics. Those differences would impact the actions associated with the different interpretations (e.g., combine expert estimates, not combine expert estimates.) The differences may result from a lack of fit of a distribution assumption on the data made in the *t*-test.

The statistical approaches were then extended to the case of group-to-group comparisons. The problem of determining whether experts are from different camps was investigated using both randomization and bootstrapping. The small population of the case studies limited group-to-group comparisons.

Future Research

There are a number of directions for future research. First, larger sample case studies could be developed to generate data for more extensive group-to-group studies. That could also facilitate a comparison of randomization and bootstrapping for investigation of knowledge bases in group-to-group studies.

Second, the generation of an exact distribution for a group-to-group approach could be extended. Rather than using constant group sizes, as in Table 6, a distribution for all possible combinations of different group sizes, m_1=1, ..., m−1 and m_2=m−1, ..., 1, respectively, could be the basis of the distribution. Such an approach allows the generation of a larger distribution size than the constant group size assumption.

Third, data from other aspects of expert systems could be investigated. This paper has focused on validation and knowledge acquisition, but other aspects, such as choosing which rules should be implemented from a multiple expert knowledge acquisition effort, might be investigated.

Fourth, this paper was concerned primarily with averaging. Future research could focus on interlocking expert judgments as discussed above. [Received: September 16, 1991. Accepted: September 10, 1992.]

REFERENCES

[1] Dungan, C. *A model of an audit judgment in the form of an expert system*. Unpublished doctoral dissertation, University of Illinois, 1983.

[2] Dungan, C., & Chandler, J. Auditor: A microcomputer-based expert system to support auditors in the field. *Expert Systems*, 1985, *2*(4), 210-221.

[3] Edgington, E. *Randomization tests*. New York: Marcel Dekker, 1980.

[4] Efron, B. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 1979, *7*(1), 1-26.

[5] Hochman, Z., & Pearson, C. J. Evaluation of knowledge-based systems in agriculture: A case study. In R. Quinlan (Ed.), *Knowledge acquisition for knowledge-based systems*. Sydney, Australia: University of Sydney, 1991.

[6] Hochman, Z., & Pearson, C. J. Evaluation of an expert systems on crossbreeding beef cattle. *Agriculture Systems*, 1991.

[7] Kaplan R. *Advanced management accounting*. Englewood Cliffs, NJ: Prentice-Hall, 1982.

[8] Lilliefors, H. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 1967, *62*, 399-402. Per the correction from the advice of C. Bates, in *Journal of the American Statistical Association*, 1969, *64*, 1702.

[9] Noreen, E. *An introduction to testing hypotheses using computer intensive methods*. New York: Wiley, 1990.

[10] Snedecor, G., & Cochran, W. *Statistical methods* (6th Ed.). Ames, IA: Iowa State University Press, 1971.

[11] Turing, A. Computing machinery and intelligence. *Mind*, 1950, *59*, 433-460.

[12] Yu, V., Fagan, L., Wraith, S., Clancey, W., Scott, A., Hannigan, J., Blum, R., Buchanan, R., & Cohen, S. Antimicrobial selection by a computer. *Journal of the American Medical Association*, 1979, *242*, 1279.

Daniel E. O'Leary is Associate Professor in the School of Business Administration, at the University of Southern California. Professor O'Leary received his Ph. D. from Case Western Reserve University, his masters from the University of Michigan, and B.S. from Bowling Green University. Professor O'Leary has published a number of papers in both decision sciences/operations research and in artificial intelligence/expert systems journals, including *Annals of Operations Research, European Journal of Operational Research, IEEE Expert, International Journal of Expert Systems: Research and Applications, International Journal of Intelligent Systems, International Journal of Man-Machine Studies* and others. Professor O'Leary is the founder of the AI in Business section of the American Association for Artificial Intelligence, on the program committee for the IEEE Conference on Artificial Intelligence Applications, and Program Chair of the 4th International Conference on Expert Systems in Accounting, Finance, and Management.