Knowledge acquisition from multiple experts: An empirical study
Daniel E O'Leary
Management Science; Aug 1998; 44, 8; ABI/INFORM Global
pg. 1049

Knowledge Acquisition from Multiple Experts: An Empirical Study

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

Expert systems often employ a weight on rules to capture conditional probabilities. For example, in classic rule-based settings, $\Pr(h|e) = x$ is used to mean "If e is known to be true then conclude h is true with probability x." Further, other probability-based approaches, such as influence diagrams and Bayes' Nets are increasingly being used to support decision making through decision support systems. Although algorithms for these systems have received substantial attention, less attention has been given to knowledge acquisition of probabilities used in these systems. However, the underlying probabilities are critical because they lead the user to particular solutions. Accordingly, the purpose of this paper is to investigate the quality of probability knowledge when it is acquired from groups or individuals.

This paper summarizes the results of an empirical cognitive study on the ability of individuals and groups to provide consistent sets of probabilities $\Pr(A)$, $\Pr(B)$, $\Pr(A|B)$ and $\Pr(B|A)$. The analysis of these probabilities allowed the study of the ability of subjects to account for Bayes' theorem reversals, a basic assumption made by virtually all algorithms. It was found that knowledge acquisition from **groups** provided more correct orderings to the probabilities than knowledge acquisition from individuals. This suggests that knowledge acquisition from groups is more likely to obtain correct probability knowledge.

(*Artificial Intelligence; Decision Support Systems; Knowledge Acquisition*)

1. Introduction

Rule-based expert systems (e.g., Buchanan and Shortliffe 1985) explicitly employ conditional probabilities in the generation of weights on rules for Bayesian diagnosis problems. In addition, increasingly, influence diagrams (e.g., Howard and Matheson 1981) and Bayes' Nets (e.g., Pearl 1988) are being used to support decision making in decision support systems (DSS) that employ artificial intelligence (AI). Both tools are based on Bayes' Theorem and classic probability. Although algorithms for these intelligent DSS have received substantial attention, less attention has been given to knowledge acquisition of probabilities (weights) used in these systems. However, knowledge acquisition of the underlying probabilities is critical because ultimately they lead the user to particular solutions. Accordingly, the purpose of this paper is to investigate knowledge acquisition of the probability estimates and what happens when groups, i.e., "multiple experts," rather than individuals are used to estimate probabilities.

As used in MYCIN (Buchanan and Shortliffe 1985, p. 79), "$P(h|e) = X$ means IF: e is known to be true THEN: conclude that h is true with probability X." Thus, it is critical that correct knowledge about probabilities be acquired. Further, Bayes' theorem reversals are used extensively by algorithms in Bayes' nets and influence diagrams. Bayes' theorem reversals refer to determining the probability of A given B (or B given A), using prior probabilities of A and B, and the probability of B given A (or A given B). Bayes' theorem reversals allow the development of a network of probabilities to a convenient form. Initial versions of the probability network might involve the probability of B given A, but it may be more convenient to work with the probability of A given B. Further, Bayes' theorem reversals, limit the

0025-1909/98/4408/1049$05.00
Copyright © 1998, Institute for Operations Research and the Management Sciences

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

probability estimates that need to be solicited in the system development. Using Bayes' theorem, an expert is asked to assess the probability of A, B and the probability of B given A, and then the probability of B and A is derived. Accordingly, Bayes' theorem reversals are dependent on the "quality" of the conditional probabilities generated through knowledge acquisition.

Unfortunately, research indicates that there may be some ambiguity regarding probability knowledge. Cognitive science researchers (e.g., Tversky and Kahneman 1983) have found that individuals exhibit fallacies in their probability reasoning. As a result, there is interest in examining the ability of knowledge acquisition approaches to ensure generation of elicited probabilities that follow the axioms of probability.

1.1. Problem Statement and Findings

The specific purpose of this paper is to analyze empirically the quality of probability knowledge acquired from individuals and groups. Probabilities are used to compare knowledge acquisition from individuals and groups because there is a "right answer," as opposed to other topics of knowledge acquisition, where there may be multiple correct answers.

Groups of subjects provided probability orderings that were consistent with Bayes' theorem, more often than did individuals. Thus, this suggests that knowledge acquisition of probability estimates from groups would appear to provide a more consistent source of probability-based knowledge than individuals. Further, individual subjects generally did not generate probability orderings of conditional probabilities consistent with Bayes' rule. Individuals, more than groups, did not appear to distinguish the probability of A given B from the probability of B given A.

1.2. This Paper

This paper proceeds as follows. Section 2 summarizes critical issues in knowledge acquisition and knowledge acquisition of probabilities from multiple experts. Section 3 provides some background on notation and Bayes' theorem. Section 4 summarizes some related findings in cognitive science and groups. Section 5 describes an experiment designed to determine the ability of individual and groups of subjects to employ Bayesian reasoning. Section 6 summarizes the findings. Section 7 investigates some approaches that can be used to try to assure quality probability estimates. Section 8 summarizes some extensions that can be made to the current research. Section 9 briefly summarizes the paper.

2. Knowledge Acquisition and Knowledge Acquisition of Probabilities Solicited from Multiple Experts

There is a substantial literature on knowledge acquisition (e.g., Holsapple and Wagner 1995). Although the primary focus has been on knowledge acquisition from single experts, other researchers have examined knowledge acquisition from multiple experts (e.g., Shpilberg et al. 1986).

2.1. Knowledge Acquisition

Knowledge acquisition plays an important role in many artificial intelligence projects, such as the well-known KADS approach developed in Europe (Schreiber et al. 1994). As a result, there have been a number of surveys regarding the various tools and approaches used in knowledge acquisition (e.g., Boose 1989, Kim and Courtney 1988).

Researchers have established various theories of the primary variables influencing knowledge acquisition. As noted in Holsapple and Wagner (1995), variables attributed as influencing knowledge acquisition include problem domain factors (e.g., Clancy 1985), organization factors (Prerau 1987) and human factors (Shpilberg 1986). Further, previous research has focused on both qualitative knowledge (e.g., "rules") and certainty factors (e.g., probabilities on the rules). Qualitative knowledge has been investigated using a number of approaches, including, e.g., empirical (e.g., Burton et al. 1987, Holsapple and Raj 1994) and field research (e.g., Watkins and O'Leary 1987). Quantitative knowledge has been the focus of only a few other researchers (e.g., Holsapple and Wu 1994, and Tonn and Goeltz 1990).

2.2. Knowledge Acquisition from Multiple Experts

Knowledge acquisition from multiple experts was proposed (Mittal and Dym 1985, p. 32) "To avoid some of the pitfalls of relying on a single expert. . . ." In particular, they criticized the ". . . simplistic view that one expert's rules constitute expertise." Some of the most interesting

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

accounts of expert systems employ knowledge acquisition from multiple experts (e.g., Shpilberg et al. 1986). In spite of the recognition of the importance of knowledge acquisition from multiple experts and its visibility, there have been limited efforts directed towards this issue (Botten et al. 1989) and the literature of multiple expert knowledge acquisition of probability-based knowledge for knowledge-based systems has received only recent attention. Knowledge about probabilities can be acquired from multiple experts in a number of ways (O'Leary 1993). First, probabilities can be acquired individually from multiple experts and then combined in a system (e.g., O'Leary 1997). Combination can be done either at the time of the model building or in the model at the time it is run (e.g., Dungan 1983). Second, rather than multiple probability estimates, a single set of probability estimates can be acquired from a group of experts. This last approach has received little attention in the artificial intelligence literature, to date. Despite the existence of these (and other) approaches to gathering probability information for knowledge-based systems, there have been no efforts to determine which approach "works best." Accordingly, this paper compares the quality of conditional probabilities generated from groups and from individuals, where quality is represented by the extent to which generated probabilities satisfy Bayes' Theorem.

3. Probability Theory: Notation, Bayes' Theorem, Soliciting Probabilities and Testing Estimates

This section provides a brief summary of notation, Bayes' theorem, soliciting conditional probabilities and a review of approaches developed to test the consistency of probability estimates generated for rule-based systems and influence diagrams.

3.1. Notation

Let $\Pr(A)$ be the probability of A. Let $\Pr(A \mid B)$ be the probability of A given B. For exposition purposes, in this paper the conditional probability $\Pr(A \mid B)$ will be assumed to be a rule-based summary of probability on "If B then A." Generally, $\Pr(A \mid B)$ is viewed as providing a "causal" structure (pointing from causes to effects), while $\Pr(B \mid A)$ provides a "diagnostic" structure (pointing from evidence to hypothesis).

3.2. Bayes' Theorem

Bayes' theorem states that

$$\Pr(A \mid B) = \Pr(B \mid A)^* \Pr(A) / \Pr(B).$$

That theorem leads us to the following "weak" version of Bayes' theorem.

THEOREM 1. *If* $\Pr(A) > (= \text{ or } <) \Pr(B)$ *then* $\Pr(A \mid B) > (= \text{ or } <) \Pr(B \mid A)$.

These results define orderings between the two pairs of probabilities. If decision makers use Bayesian reversal then they will order probabilities according to the results in theorem 1. Whether or not decision makers (students or professionals or groups) order probabilities according to theorem 1 is an empirical question.

3.3. Knowledge Acquisition of Probabilities

Conditional probabilities can be assessed using either a causal structure ($\Pr(A \mid B)$) or a diagnostic structure ($\Pr(B \mid A)$) or both. The empirical findings seem to substantiate using the causal form as a preferred approach. Tversky and Kahneman (1980) found that subjects apparently prefer to encode experience in causal schema. As a result, as noted by Pearl (1988, p. 151) ". . . rules expressed in causal form are assessed more reliably."

Unfortunately, individual differences can result in one expert providing better estimates based on diagnosis probabilities. Further, reasoning in some domains seems to be consistent with diagnosis (e.g., medical domain and most "trouble shooting problems"). As a result, just soliciting the causal or just soliciting the diagnostic probabilities may not result in correct probabilities, or may not be consistent with the expert's ability to estimate probabilities.

These differences are illustrated in the structure of the mediums used to model processes. In the medical expert system MYCIN (Buchanan and Shortliffe 1985) rules point from evidence to hypothesis, for example, symptom to disease. While, with Bayes' nets and influence diagrams, there are arcs from causes to effects and from conditions to consequences (Pearl 1988, p. 151).

4. Background: Cognitive Science and Group Theory

This section provides a brief background into what has been called "human fallacy" research, because of its

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

apparent concern for finding errors made by humans. In addition, this section provides a brief discussion of group theory.

4.1. Cognitive Science

Cognitive science researchers (e.g., Tversky and Kahneman 1983) have noted that individual subjects appear to exhibit probability fallacy in their reasoning. They provided evidence that people assess the probability of the intersection of two events to be greater than the probability of at least one event. They have labeled this as the conjunction fallacy.

As an example, Tversky and Kahneman (1983) conducted the study labeled "Predicting Wimbledon." That research asked subjects to rank the probability of four different outcomes:

(a) Borg will win the match;
(b) Borg will lose the first set;
(c) Borg will lose the first set but lose the match;
(d) Borg will win the first set but lose the match.

They found on average that the probability assigned to (c) was greater than the probability assigned to (b). Because this is inconsistent with probability theory, it was referred to as the "conjunction fallacy."

The conjunction fallacy has little direct interest to reasoning under uncertainty in AI systems or influence diagrams. However, the existence of the conjunction fallacy indicates that it is likely that subjects do not reason in a manner consistent with Bayes' theorem, because conditional probability is a function of conjunction probability.

In research that is more directly related to the estimation of conditional probabilities, Tversky and Kahneman (1980) found that ". . . there may be pairs of political events for which $\Pr(A)$ is judged greater than $\Pr(B)$, but $\Pr(A \mid B)$ is judged less than $\Pr(B \mid A)$." This is a violation of probability theory.

In a later study, Einhorn and Hogarth (1986) found that ambiguity in the temporal sequence of the events A and B, can influence the estimation of $\Pr(A \mid B)$ and $\Pr(B \mid A)$. For example, consider the case where there is a test for cancer, such as a mammography. A test will be found to be positive, only after the patient has cancer. In that situation, the test predicts the state that existed prior to the test. Thus, the sequence goes from finding the effect to determining the cause. Accordingly, they note (Einhorn and Hogarth 1986, p. 9) that people have difficulty distinguishing between $\Pr(A \mid B)$ and $\Pr(B \mid A)$.

Thus, it appears that individuals do not reason probabilistically in a manner that is consistent in with the weak version of Bayes' Theorem. As a result, we have the following hypothesis.

HYPOTHESIS 1. *Subjects will not reason probabilistically in a manner that is qualitatively consistent with Bayes' Theorem.*

4.2. Groups vs. Individuals

Throughout the literature that concentrates on group behavior, there is the notion that group behavior is different than individual behavior. As noted by Weick (1969, p. 32), "People in aggregates behave differently than do people in isolation." In addition, Nunamaker et al. (1993, pp. 127–128) argue that "certain aspects of (groups) . . . improve outcomes (process gains), whereas others impair outcomes (process losses) relative to the efforts of the same individuals working by themselves or those of groups that do not experience them. . . ." They go on to establish a number of sources of group process gains (e.g., synergy and learning) and sources of group process losses (e.g., conformance pressure and free riding). As they note, an important source of gains is the increased information and knowledge available in groups compared to individuals. Thus, although individuals seem to execute fallacies in reasoning it is not clear that groups would use the same fallacious reasoning. As a result, knowledge acquisition for probabilistic systems might be more effective using groups as the source of knowledge for intelligent DSS.

One variable of group behavior of particular importance is the size of the group, particularly in the case of small groups (Weick 1969, Simmel 1950). Apparently the crucial transitions in group sizes are from one to two persons, from two to three persons, from three to four persons, from four to seven persons and from seven to nine persons (Weick 1969). In particular, Weick (1969, p. 38) refers to triads as the basic unit of analysis of groups. The triad is particularly important because it is the smallest group size that allows for alliance of two group members against one. The alliance allows for cooperation, control and competition. In any group of size four or more, a group size of three can be a subgroup. Because the triad is the most critical group size, the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

groups in this research were developed with three members each.

4.3. Perceptual Sets

The notion of perceptual sets argues that people see what they are attuned to see, based on past experience. Subjects carry their perceptual sets from situation to situation. As a result, if subjects have had training in probability then we would expect them to carry that experience with them as part of their perceptual set.

AI and cognitive science researchers have made frequent use of perceptual sets (e.g., Simon 1969). For example, expert system developers generally assume that the knowledge acquisition task is to acquire as much of the expert's perceptual set as is feasible and necessary, in order to build the specific system.

This notion of a perceptual set is useful in differentiating a group from an individual: the group's perceptual set is the union of the perceptual sets of the individuals. Thus, given similar training in probability, we would expect that groups of individuals would more likely include the appropriate probability theory than any single individual. This leads us to the following hypothesis:

HYPOTHESIS 2. *Groups reason differently (and more correctly) about probabilities than individuals.*

5. Empirical Study

An empirical study was made to investigate the hypotheses and to determine whether individual and group decision makers follow the ordering of probabilities suggested by Bayes' theorem and summarized in Theorem 1.

5.1. Test Instrument

The test instrument was one that derived from the Tversky and Kahneman (1983) approach to "Predicting Wimbledon." The subjects were asked to rank the probabilities of set of events that included two events A and B, and the structuring of those two events as "if A then B" and "if B then A." The instrument is summarized in Table 1.

Table 1 Case Study

You are planning a review of Electra's internal controls. Although the company has not emphasized a strong network of detailed control procedures, top management closely monitors the operations and overall management controls generally serve as an adequate substitute for detailed controls. Please rank order the following outcomes from most to least likely (1 = most likely, 2 = second most likely, etc.) If any outcome is as likely as another, please indicate so by using the same number.

	Ranking
A. The system of internal controls is strong	___
B. Initial compliance testing reveals some errors	___
. . .	
F. If initial compliance testing reveals some errors then the system of internal controls is strong	___
G. If the system of internal controls is strong then initial compliance testing reveals some errors	___
. . .	
I. Initial compliance testing reveals some errors but the system of internal controls is strong	___

5.2. Subjects

There were three groups of subjects. First, 45 business consultants (P) participated as subjects. These subjects averaged about 2.5 years of experience. Second, students were used to form groups. Thirty groups (G) of three members each, averaging about three years of professional business experience, completed the test instrument as an in-class exercise for group project. Prior to this in-class assignment, the groups had completed three group projects. Thus, these groups were actually working groups, analogous to those that might be seen in general organizational environments. Third, 31 individual students (S), with 2.5 years of experience completed the test instrument as an in-class exercise.

Each set of subjects was administered the instrument in their normal settings. The consultants were administered the instrument at their offices and the students were administered the instrument on campus. Each group could submit only one set of rankings. There were no time constraints and there was no promulgated methodology. If there was group conflict the groups dealt with it in a manner that met their needs.

5.3. Data Analysis

The data analysis used two different approaches. First, summary statistics regarding the rankings of the probabilities of sets of events were analyzed. The two sets of probabilities of events ($\Pr(A)$ and $\Pr(B)$, and $\Pr(A \mid B)$ and $\Pr(B \mid A)$) were each analyzed to determine if the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

probabilities orderings $\Pr(A)$ and $\Pr(B)$, and $\Pr(A \mid B)$ and $\Pr(B \mid A)$ came from the same distributions. Because Bayes' Theorem is an equality, it is expected that if the means of the distributions of $\Pr(A)$ and $\Pr(B)$ are from different (the same) distributions, then the $\Pr(A \mid B)$ and $\Pr(B \mid A)$ would be from different (the same) distributions, and conversely.

Second, individual and group responses were investigated in order to determine if groups generated a higher proportion of correct responses, i.e., fewer probability ordering "violations." The appearance of a violation in ordering can be found by inspecting the order of the probabilities of the sets of events by each of the subjects and groups of subjects. The number of violations for each group of subjects were then compared to determine which set of subjects performed the best and if there were statistically significant differences between individual and group performance.

6. Findings

The results substantiate the hypotheses. Individuals had average rankings that apparently differentiated between $\Pr(A)$ and $\Pr(B)$, but not between $\Pr(A \mid B)$ and $\Pr(B \mid A)$. In addition, individuals had more violations of Bayes' Theorem (H1). Groups had a statistically significantly different proportion of correct responses, and they had a higher percentage of correct responses, than individuals. In this case, knowledge acquisition from groups is more likely to result in probabilities that are consistent with Bayes' Theorem.

6.1. Means and Standard Deviations Come from Different Distributions

In Table 2, it can be seen that the means of the orderings of the probabilities of the two events (A and B) appear to come from different distributions (at the 0.01 level) for each of the different samples. However, we cannot reject the hypothesis that the means of the orderings of the two conditional sets ($\Pr(A \mid B)$ and $\Pr(B \mid A)$) come from the same distribution for each of the different samples for individuals and groups. In contrast, the group means for the conditional probabilities are statistically different at the 0.07 level.

In addition, the variances of the orderings of the probabilities of the two events (A and B) appear to come from different distributions (at the 0.005 level or better).

Table 2 Summary Statistics for Events(1, 2, 3)

Event	Subjects	Mean	Standard Deviation
$\Pr(A)$	P (45)	2.75@	1.479@
	S (31)	3.70&	2.193&
	G (30)	3.17^	1.234^
$\Pr(B)$	P	1.35@	0.857@
	S	2.00&	1.00&
	G	1.70^	0.534^
$\Pr(A \mid B)$	P	3.35	1.667
	S	5.19	1.833
	G	4.83 #	1.858
$\Pr(B \mid A)$	P	3.68	1.819
	S	5.29	2.100
	G	4.30 #	1.822

Notes.

1. Based on the rankings of the probabilities of events. A was event 2 and B was event 1.
2. P = Professionals; G = Groups, S = Individuals.
3. The test that the means of rankings for $\Pr(A)$ and $\Pr(B)$ are different is statistically significant at the 0.01 level usnig a t-test fro each of the pairs @, &, * and ^ at the 0.005 level or better for the standard deviations. The test that the means or the standard deviations of the rankings of $\Pr(A \mid B)$ and $\Pr(B \mid A)$ are different was not statistically significant for professionals or individuals but was significant at the # = 0.07 level for groups.

However, we cannot the reject the hypothesis that the variances of the orderings of the conditional probabilities come from the same distribution.

Thus, on average, individual subjects differentiate between the ordering of the rankings of the probabilities of basic events but not the conditionals. This is in contrast to the equality nature of Bayes' theorem which suggests, e.g., if the means for one pair are different then the means for the other pair would also be different. On the other hand, on average, groups differentiate between both rankings of probabilities of basic events and the conditionals. As a result, group behavior appears more consistent with Bayes theorem.

6.2. Violations of Orderings by Subjects

An alternative approach to examining summary statistics about the rankings is to examine actual violations in rankings of probabilities of events. The results are summarized in Table 3. Using a test of proportions (Dixon and Massey 1969) it was found that the number of violations are statistically significantly different for individuals and groups. A "z-test" of different propor-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 3 Violations Analysis

Table 3A. Types of Outcomes(1)

	P	S	G	Total
Violations(2)	30	16	9	55
Correct	15	15	21	51
Total	45	31	30	106

Notes.

1. P = Professionals; S = Students; G = Groups
2. A violation refers to a violation of Theorem 1.

Table 3B. Test of Proportions[1]

	Professionals	Students	Group
Professionals	—		
Students	1.319	—	
	(0.10)		
Groups	3.11	1.712	—
	(0.001)	(0.05)	

[1] Top number is the z-score when computed using the test of proportions (Dixon and Massey 1969) and bottom number is statistical significance. Table is symmetric so only bottom is shown.

tions was used to statistically test the differences. It was found that the difference between the proportion of violations for professionals and groups were significantly different at the 0.001 level, and the difference between students and groups were statistically significantly different at the 0.05 level. Groups provided a statistically significantly different number of ranking violations than both groups of individuals.

There were two possible categories used to assess the quality of the orderings: violation and correct. A "violation" is a violation of Theorem 1. If there is no violation of Theorem 1 then the order is correct. Table 2 suggests that it would be unlikely that $\Pr(A)$ and $\Pr(B)$ would be ranked as equal by any group. Based on that result and Bayes' Theorem, that would suggest that it is also unlikely that $\Pr(A \mid B)$ and $\Pr(B \mid A)$ would be equal. Although the average attributed to the conditional probabilities by individuals are not statistically different, the difference between the probabilities of the events is statistically different and thus, this is in contradiction to Bayes' Theorem.

The groups had the best performance, substantiating Hypothesis 2. Less than one-third of the groups had performance that was inconsistent with Bayes' theorem.

6.3. Discussion

The individual subjects did not effectively differentiate between causal and diagnostic versions of the conditional probabilities, on average. It is difficult to argue that they developed one correctly and the other incorrectly, because if they had one correct, it would seem that it would have been ranked significantly different than the one they had incorrect. It appears that when confronted with both causal and diagnostic versions, that the subjects are unable to differentiate between them, because they ranked them the same.

The findings may have been influenced by the nature of the task. The temporal sequence of the events is not the same as the causation. It is by the "initial compliance testing" that we determine the status of the system. However, "a strong system of internal controls" is what causes finding only some errors. This is analogous to the discussion of analysis of cancer, in terms of cause and effect, by Einhorn and Hogarth (1986). Such concerns need to be addressed in future research.

7. Implications

The results of this study have implications for development of decision support systems that account for uncertainty and for influence diagrams. Bayesian reversals may be quite correct mathematically (i.e., normatively), but don't appear to model subject's reasoning. Although groups tend to improve the consistency of the estimates, subjects generally don't effectively differentiate between causal and diagnostic conditional probabilities. As a result, construction of probability-based DSS needs to account for this potential inconsistency. There are a number of potential solutions to facilitate system development.

Training in Probabilities. This research finds that both individual students and consultants have difficulty differentiating between $\Pr(B \mid A)$ and $\Pr(A \mid B)$, however, the student performance was better than the consultants. One explanation is that the students training in conditional probabilities was more recent. As a result, training of domain experts probably is an important tool in ensuring that conditional probability estimates are correct.

Use Only One of $\Pr(A \mid B)$ or $\Pr(B \mid A)$. One approach to limiting the impact of the lack of consistency

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

between rankings of probabilities is to choose one of $\Pr(A \mid B)$ or $\Pr(B \mid A)$ that the expert is most likely to get right. Unfortunately, as noted earlier, it is unclear if diagnostic or causal versions are most effective. Further, unless they are periodically gathered and contrasted, system developers will not know if experts on whom probabilities are based are differentiating between the two conditional probabilities.

GATHER ADDITIONAL INFORMATION. This can be done by forcing the collection of additional information. In particular, the information gathering requirements could include $\Pr(A)$, $\Pr(B)$, $\Pr(A \mid B)$ and $\Pr(B \mid A)$. If those estimates are not in concert with Theorem 1 or Bayes' theorem then additional work is required so that decision maker comes to a set of estimates that are consistent. Unfortunately, such a check has a very high cost. The results can suggest that those probabilities must be gathered for all pairs, triplets, . . . , etc. Effectively, the system cannot perform arbitrary Bayesian reversals without actually generating and checking the probabilities (if correct probabilities are desired). This approach eliminates one of the primary computational benefits of influence diagrams, Bayesian reversals.

AUTOMATE ACQUISITION OF PROBABILITIES. Another approach would be to get the human out of the process by automating the acquisition of probability estimates. Unfortunately, this approach limits taking advantage of human expertise.

IN ANY CASE: USE EXTREME CARE. The results indicate that extreme care must be used in the generation of probabilities for influence diagrams and other probability based forms of analysis. This indicates that as feasible, various data checks need to be incorporated in systems designed to solicit probabilities. Checking the consistency of probabilities elicited for use in the system has not received attention in the literature or in the construction of software tools for probability reasoning. In large measure, this is likely because of the assumption of the ability to perform Bayesian reversals.

GENERATE SAMPLES OF PROBABILITIES. Periodically, it may be beneficial to generate an entire set of probabilities, such as $\Pr(A)$, $\Pr(B)$, $\Pr(A \mid B)$ and $\Pr(B \mid A)$, in order to test the quality of the numbers that are being generated. If discrepancies are found, then biases of the expert can be used to help generate better estimates.

8. Extensions[1]

This paper provides an initial analysis of the issue of group ranking of conditional probabilities. It shows that in some problems groups provide better probability estimates than individuals.

8.1. Group Operation

This research did not gather any formal statistics regarding how groups solved the problem. However, I observed informally that in a number of groups, individuals realized that probability theory could help solve the problem. A number of the test instruments showed probability formulas or examples. As a result, in many cases, it could be observed that the group members brought their knowledge of probability to the solution of the problem. Future research could focus explicitly on attributes of these group processes.

8.2. Perceptual Set Synergism and Dissonance

This research is based on the group perceptual set being a union of the individual perceptual sets, and does not consider the particular source of improvement. However, in other settings groups have generated dissonance as exhibited through such dysfunctional activities as "group think," where group members get focused a particular view and ignore large amounts of data (e.g., Schlesinger 1965). In addition, groups are sometimes thought to bring synergy to projects (e.g., Jessup and Valacich 1993). This research could be extended by focusing on those possible circumstances that could change the quality of the results deriving from synergy and dissonance. Such research could be integrated with the analysis of group processes as discussed in §8.1.

8.3. Certainty Factors

MYCIN certainty factors (Buchanan and Shortliffe 1985) were developed using Bayesian probability theory and certain assumptions to limit the computational and knowledge requirements. Typically, these uncertainty factors can be constructed using one of two different approaches. First, probabilities can be solicited and then using those probabilities, uncertainty factors can be constructed. Second, the uncertainty factors themselves can

[1] I would like to thank the referees for bringing these extensions to my attention.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

be solicited directly. In the first case, the certainty factors are computed using the conditional probability of the hypothesis, given the evidence (Buchanan and Shortliffe 1985, p. 248). As a result, the quality of the estimated probabilities is likely to be influenced by the findings in this paper. On the other hand, if the probabilities are neglected and the factors estimated directly then the apparent dependence on conditional probabilities can be avoided. However, there is substantial evidence (e.g., O'Leary 1990) that when this kind of approach is used that the estimates are likely to violate the underlying probability structure on which the rules are based. A number of researchers (e.g., Holsapple and Wu 1994) have found that certainty factors are not a "good fit" with most users. The direct or indirect estimation of conditional probabilities may be among the underlying reasons for their findings.

8.4. Fuzzy Sets

Fuzzy set approaches (Zadeh 1965) can be integrated with probability measures (Zadeh 1968). For example, O'Leary (1994) investigates an integrated fuzzy set and probability-based approach in expert systems to model the validation process. That analysis integrates classic Bayesian probabilities and fuzzy sets. As a result, in that model, the quality of the conditional probabilities ultimately influences the quality of results. Accordingly, for those situations where probability measures are used with fuzzy set models, the results here would apply, i.e., we would expect the group generated results to be better.

In addition, groups may be able to generate better fuzzy set measures, but that could be difficult to determine. In probability theory there is a theory that tells how things should be ordered so we can determine if groups or individuals ranked things they way they were supposed to be. Unfortunately, unlike probability theory, there is seldom a constraining fuzzy set theory that would set expectations for experimental purposes.

8.5. Subjects

Working groups of students, all with business experience, were compared to consultants with roughly the same amount of experience and to other students. Either group provides a direct comparison basis, and groups were able to generate better results than either pool of individuals. This research focused on groups with experience and my subject pool for experienced groups was limited to students. However, researchers with access to other subject pools could examine other groups as sources.

8.6. Scope of the Research

Although this research establishes an initial set of findings it also raises other issues that can lead to future research. In addition to the topics discussed in this section, the results developed here for a problem in a business setting could be extended to other problem settings and problem types, particularly those settings where groups play an important role. This research shows that in some settings groups provide the ability to rank probability information consistently with Bayes' Theorem more frequently than individuals. The question remains for future research to determine what happens in other settings.

9. Summary and Applications

This final section summarizes the paper and provides some additional applications of the research.

9.1. Summary

This paper presents an empirical study that indicates that knowledge acquisition of probabilities from groups is more likely to generate consistent probability knowledge than knowledge acquisition from individuals. In addition, knowledge acquisition of conditional probabilities, from professional decision makers is likely to result in probability orderings not consistent with the underlying theory. Individuals, and to a lesser extent, groups, do not appear to differentiate between different conditional probabilities.

9.2. Applications

There are a number of implications of this cognitive study for decision support systems. In particular, the results provided are critical to those who use probability-based systems, generate probabilities for those systems or build shells for probability use with influence diagrams, etc.

- Knowledge acquisition from groups of individuals appears more likely to generate probability orderings consistent with the underlying probability theory.
- Those who use probability-based systems need to know that model probabilities are consistent with Bayes' Theorem. If the probabilities are not appropriate then the system is likely to lead the user to some solutions that are not optimal.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

- Those who develop decision support systems, and accordingly generate probabilities, need to be cognizant of the potential for violation of Bayes' Theorem. These results suggest that substantial care be used in the generation of probabilities. At the very least, there should be periodic sampling of pairs of probabilities to ensure that consistent probabilities are gathered.
- Those who build probability-based shells used by decision support developers should make sure that tests of consistency are available. In addition, those developers should also suggest to shell users that Bayesian reversals may result in probabilities estimates that are not correct.[2]

[2] The author would like to thank the anonymous referees and Associate Editor for their extensive comments on two earlier versions. In addition, a previous version of this paper was presented at the IJCAI-95 "Workshop on Building Probabilistic Networks." The author would like to acknowledge the comments of the three anonymous referees for that version of this paper prepared for that workshop.

References

Boose, J. 1989. A Survey of Knowledge Acquisition Techniques and Tools. *Knowledge Acquisition*. **1** 3–38.

Botten, N., A. Kusiak, T. Raz 1989. Knowledge Bases: Integration, Verification and Partitioning. *European Journal of Operational Research*. **42** 111–128.

Buchanan, B., E. Shortliffe 1985. *Rule-Based Expert Systems*. Addison-Wesley, Reading, MA.

Burton, A., N. Shadbolt, A. Hedgecock, G. Rugg 1987. A Formal Evaluation of Knowledge Elicitation Techniques for Expert Systems. *Proceedings of the First European Workshop on Knowledge Acquisition for Knowledge-based Systems*. September.

Clancy, W. J. 1985. Heuristic Classification. *Artificial Intelligence*. **27** 298–350.

Dixon, W., F. Massey 1969. *Introduction to Statistical Analysis*. McGraw-Hill, New York.

Dungan, C. 1983. A model of audit judgment in the form of an expert system. Unpublished Ph.D. dissertation, University of Illinois.

Einhorn, H., R. Hogarth 1986. Judging Probable Cause. *Psychological Bulletin*. **99** 3–19.

Holsapple, C., V. Raj 1994. An Exploratory Study of Two KA Methods. *Expert Systems*. **11** 77–87.

——, W. Wagner 1995. Contextual Factors in Knowledge Acquisition. *International Journal of Expert Systems*. **8** 327–348.

——, J. Wu 1994. Certainty Factor Algebras: Comparing Conventional Mappings and Experimental Results. *International Journal of Intelligent Systems in Accounting, Finance and Management*. **3** 35–46.

Howard, R., J. Matheson 1981. Influence Diagrams. *The Principles and Applications of Decision Analysis*. Volume II. Strategic Decisions Group, Menlo Park, CA, 720–762.

Jessup, L., J. Valacich 1993. *Group Support Systems*, Macmillan, New York.

Kim, J., J. Courtney 1988. A Survey of Knowledge Acquisition Techniques and the Relevance to Managerial Problem Domains. *Decision Support Systems*. **4** 269–284.

Mittal, S., C. Dym 1985. Knowledge Acquisition From Multiple Experts. *AI Magazine*. Summer.

Nunamaker, J., A. Dennis, J. Valacich, D. Vogel, J. George 1993. Group Support Systems Research: Experience from the Lab and Field. L. Jessup, J. Valacich, eds. *Group Support Systems*. Macmillan, New York. 125–145.

O'Leary, D. 1990. Soliciting Weights or Probabilities from Experts for Rule-based Expert Systems. *International Journal of Man-Machine Studies*. **32** 293–301.

——, 1993. Determining Differences in Expert Judgment: Implications for Knowledge Acquisition and Validation. *Decision Sciences*. **24** 395–407.

——, 1997. Validation of Computational Models Based on Multiple Heterogeneous Knowledge Sources. *Computational and Math. Organization Theory*. **3** 75–90.

—— 1994, A Fuzzy-Set Approach for Analyzing the Validation of Expert Systems in a Multiple Agent Environment. *Expert Systems with Applications: An International Journal*. **7** 169–174.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman Publishing, San Mateo, CA.

Prerau, D. 1987. Knowledge Acquisition in Development of a Large Expert System. *AI Magazine*. **8** 43–52.

Schlesinger, A. 1965. *A Thousand Days*. Houghton Mufflin, Boston.

Schreiber, G., B. Wielinga, R. De Hoog, H. Akkermans, W. Van de Velde 1994. CommonKADS: A Comprehensive Methodology for KBS Development. *IEEE Expert*. December. 28–37.

Shpilberg, D., L. Graham, H. Schatz 1986. ExperTAX: An Expert System for Corporate Tax Planning. *Expert Systems*. **3** 53–65.

Simmel, G. 1950. *The Sociology of Georg Simmel*. K. Wolfe, ed. Free Press, New York.

Simon, H. 1969. *The Sciences of the Artificial*. 2d ed. MIT Press, Cambridge, MA.

Tonn, B., R. Goeltz 1990. Psychological Judgement in Artificial Intelligence and Expert Systems. *Expert Systems*. **7** 94–100.

Tversky, A., D. Kahneman 1980. Causal Schemata in Judgments Under Uncertainty. *Progress in Social Psychology*. M. Fishbein, ed. Lawrence Erlbaum, Hillsdale, NJ.

——, —— 1983. Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*. **90** 293–315.

Watkins, P., D. O'Leary 1987. Knowledge Acquisition for Small Scale Expert Systems from Consulting Experts: A Field Study Approach. *Proceedings of the First European Workshop on Knowledge Acquisition for Knowledge-Based Systems*. Reading Univ., Reading, England.

Weick, K. 1969. *The Social Psychology of Organizing*, Addison-Wesley, Reading, MA.

Zadeh, L. 1965. Fuzzy Sets. *Information and Control*. **8** 338–353.

—— 1968. Probability Measures of Fuzzy Events. *Journal of Mathematical Analysis and Applications*. **23** 421–427.

Accepted by John C. Henderson, former Departmental Editor; received April 23, 1996. This paper has been with the author 6 months for 2 revisions.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.